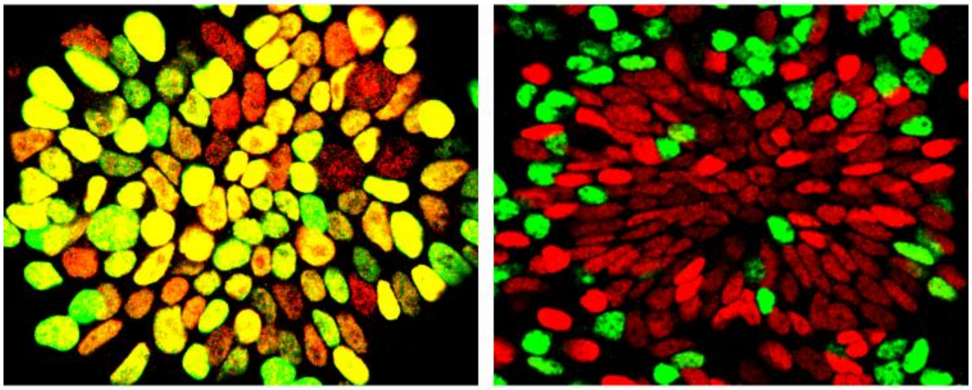
Differential Expression Analysis and Functional Interpretation of Transcriptome Changes in Neural Progenitors under Temporal Switch

Statistical and Network Biology approaches to deal with RNA-Seq data



Master of Science Thesis in the Master Programme Bioinformatics and Systems Biology

ASHWINI PRIYA JEGGARI

Department of Mathematical Sciences
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden, 2013

Differential Expression Analysis and Functional Interpretation of Transcriptome Changes in Neural Progenitors under Temporal Switch

Statistical and Network Biology approaches to deal with RNA-Seq data

By,
ASHWINI PRIYA JEGGARI

Department of Mathematical Sciences
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden, 2013

Differential Expression Analysis and Functional Interpretation of Transcriptome Changes in Neural Progenitors under Temporal Switch

Statistical and Network Biology approaches to deal with RNA-Seq data

Master of Science Thesis in the Master's Programme Bioinformatics and Systems Biology

ASHWINI PRIYA JEGGARI

Examensarbete/Institutionen för matematiska vetenskaper,
Chalmers tekniska högskola 2013:

Department of Mathematical Sciences
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
SE-412 96 GÖTEBORG
SWEDEN
Telephone: +46 (0) 31-772 1000

Department of Cell and Molecular Biology
Karolinska Institute
Box 285, SE-171 77 STOCKHOLM
SWEDEN
Telephone: +46 (0) 8 524 80000

Cover:
Figure for “Expression of neural progenitors” used with permission from Zhanna Alekseenko, Cell and Molecular Biology, Karolinska Institute

ABSTRACT

This master thesis work is dedicated to gene expression analysis in order to identify genes involved in temporal cell fate determination in the developing hindbrain. An embryonic stem cell-based differentiation protocol was used for the derivation of neural stem cells competent to produce hindbrain specific neuronal cell types: visceral motor neurons and serotonergic neurons. mRNA isolated from the neural progenitors was subjected to deep sequencing. The experiments were carried out at Cell and Molecular Biology department, Karolinska Institute, and Science for Life Laboratory, Stockholm. This work is based on the concepts of developmental and stem-cell biology, statistical and computational analysis.

An important mechanism in the emergence of neuronal diversity of central nervous system is temporal patterning, i.e. time-ordered generation of distinct cell types from a common pool of progenitors. In the hindbrain, a progenitor domain that expresses the transcription factor Nkx2.2 sequentially gives rise to visceral motor neurons (VMNs), serotonergic neurons (5HTNs) and oligodendrocyte precursor cells (OLPs). Understanding the molecular mechanism that underlies the temporal changes in the progenitor competence of Nkx2.2 progenitor cells has a potential clinical value since these neural cells are associated with neurodegenerative or neurological disorders.

This study is based on the analysis of gene expression changes during VMN, 5HTN and OLP generation from mouse embryonic stem cells in a series of transcriptome snapshots after experimental perturbations, such as gene knock-outs and blocking of the TGFβ2 signaling. The differential expression analysis on RNA-seq data raised an additional challenge since the read-based expression values, unlike those from microarrays, are not subject to traditional statistical methods (such as t-test or linear models). Further, to characterize the molecular mechanisms behind the transcriptome response the method of Network Enrichment Analysis (NEA) was applied. Here, we report genes and signaling pathways involved in the temporal patterning in the embryonic hindbrain.

Key words: Sonic hedgehog, visceral motor neurons, serotonergic neurons, oligodendrocyte precursors, RNA-Seq, statistical analysis, Network Enrichment Analysis.

PREFACE

My passion to work in a place where bioinformatics merges with medicine had been fulfilled by getting involved in a collaborated project between Science for Life Laboratory and Cell and Molecular Biology department, Karolinska Institute. It was an opportunity for me to practise bioinformatics techniques in wet-lab environment and I gained an immense knowledge that is useful for my future research.

I would like to express my gratitude to my supervisor **Andrey Alexeyenko**, who gave me a chance to get involved in this remarkable project. I am extremely thankful to him for helping me throughout the project with valuable ideas and excellent supervision that aided me to carry out the project successfully. I also want to thank **Zhanna Alekseenko and Jose Dias** for helping me with biological concepts and making me understand about neuronal stem cell concepts and answering my doubts with patience. Thank you for showing wet-lab experiments which gave me ideas beyond my bioinformatics work. I would like to thank, **Johan Ericson** for giving me the opportunity to be a part of this project in Karolinska Institute. I would also like to thank all group members from Ericson's lab for their nice discussions.

I would like to thank **Olle Nerman,** for helping me with suggestions and advices during my master's programme and being examiner for my thesis project. Also, I thank all my professors who helped me mastering in Bioinformatics and Systems Biology. All the lab courses and theoretical knowledge helped me a lot while carrying out my Master's thesis project.

I also thank all my friends from Chalmers University and University of Gothenburg for giving me good memories all through my master's years in Gothenburg. All you people will hold a special place in my heart.

Finally I thank my family, my parents **Devendar Jeggari, Indira Mamidi** who encouraged me for higher studies and supported me at every aspect of my life and my siblings **Anusha** and **Vamshi** who makes me to forget the feel that I am far away home and holding me every day of my life.

Stockholm, April 2013
Ashwini Priya Jeggari

ABBREVIATIONS

CNS	Central Nervous System
VMN	Visceral Motor Neuron
5HTN	Serotonergic Neuron
OLP	Oligodendrocyte precursor
5-HT	Serotonin
Nkx2.2	NK transcription factor related, locus2
Phox2b	paired-like homeodomain protein 2b
Foxa2	Fork head transcription factor
Shh	Sonic hedge-hog
BMP	Bone Morphogenic Proteins
RA	Retinoic Acid
wnt	wingless-related MMTV integration site
HD	Homeo-domain
NT	Notochord
FP	Floor-plate
ALS	Amyotrophic Lateral Sclerosis
ESC	Embryonic Stem Cells
Tgfβ	Transforming growth factor beta
RNA-seq	RNA-sequencing
DESeq	Differential Expression analysis for Sequence count data
EdgeR	Empirical analysis of digital gene expression in R
FDR	False Discovery Rate
AGS	Altered Gene Sets
FGS	Functional Gene Sets
NEA	Network Enrichment Analysis
GSEA	Gene Set Enrichment Analysis
GO	Gene Ontology terms
KEGG	Kyoto Encyclopedia of Genes and Genomes
WT	Wild-type

CONTENTS

1 INTRODUCTION

The central nervous system is a major control center of the body which allows an individual to perform many tasks from sensory perception and motor coordination to behavior and memory. All these functions are being carried out by hundreds of functionally distinct neuronal subtypes that establish specific synaptic connections with other neurons. In addition to neurons, the adult brain contains astrocytes and oligodendrocytes together called macroglial/glial cells, which provide supportive functions to neurons. During vertebrate embryonic development, neurons are generated from pools of progenitor cells. In the ventral hindbrain, a pool of progenitor cells that express the transcription factor Nkx2.2$^+$ generates in a defined temporal order visceral motor neurons, serotonergic neurons and oligodendrocyte precursors [1,2]. Uncovering the molecular mechanisms that regulate temporal changes in progenitor competence of Nkx2.2$^+$ progenitor cells have a clinical value. These hindbrain neural subtypes have been associated with severe neurodegenerative and neurological disorders. For example, serotonergic neurons are involved in the synthesis of the neurotransmitter serotonin and alterations in serotonergic neuron function have been related to neurodevelopment disorders such as depression, anxiety, autism, disorders of energy balance, schizophrenia, and sudden infant death syndrome. Visceral motor neurons are involved in the activity of smooth muscle fibers, cardiac muscles and glands. VMNs degeneration results in the muscle movement disorders such as amyotrophic lateral sclerosis, primary lateral sclerosis and progressive muscular atrophy [3,4,5,6].

Embryonic stem cells are a powerful tool to study neural development where CNS progenitor cells can be generated from ES cells and from these cells it is possible to generate the neurons in the presence of specific mitogen and signaling molecules [7]. Identification and selection of differentially expressed genes from the RNA-seq data is quite challenging. As the data is of count nature, which is different from the micro-array intensity data, the traditional methods inherited from the micro-array toolbox could not be applied directly. The methods such as DESeq, Edge R used to analyze the RNA-seq data were not reliable to current data, as the samples do not contain biological replicates. This analysis uses bioinformatics, statistical and systems biology approaches to carefully evaluate and select the methods that are more relevant to the situation. Further in order to understand the molecular mechanism behind the transcriptional response that enables the temporal switch, enrichment analysis was performed to access the relationship between differentially expressed genes and known gene-sets representing biological pathways. Enrichment analysis was carried by Gene Set Enrichment Analysis and Network Enrichment Analysis. GSEA gives a list of ranked genes from the Altered Gene Sets that are over represented in Functional Gene Sets. Moreover, GSEA ignores the functional relations between AGS genes themselves and between AGS and outside pathways. Whereas NEA, uses all available network links scattered over the network to test enrichment hypothesis of functional associations between AGS and FGS.

In the following, I describe the method we followed to analyze the RNA-seq data, statistical tests that were implemented to calculate the differential attributes such as p-adjusted (q-values) and fold-change values, cross-comparative analysis between wild-type and mutant samples used for the identification of the novel candidate genes associated with VMN-to-5HTN cell fate switch, visualization of identified differential expressed (DE) genes and description of functional relations by NEA.

1.1 AIMS AND OBJECTIVES

The aim of the work presented in this master thesis was to understand the concrete mechanisms that underlie the switch in a series of global transcriptome measurements using wild-type differentiated ES cells or mutant ES cells for key transcription factors which are essential for VMN-to-5HTN fate switch.

Specific aims

1. To investigate the method to deal with RNA-seq data in the absence of biological replicates and to obtain robust and statistically significant estimates of gene differential expression for both high and low abundant transcripts.

2. To perform higher level cross-comparisons between different pairwise contrasts of wild-type and mutant samples to identify novel genes associated with cell-fate switch and 5HTN generation.

3. To evaluate applicability and performance of different methods of enrichment analysis, to analyze the lists of DE genes, to interpret how these genes are involved in biologically relevant functional pathways and processes.

2 BACKGROUND

During the early stages of CNS development, individual progenitor cells acquire distinct properties in the accordance to their spatial positions along the anterior/posterior (A/P) and dorsal/ventral (D/V) axes of the neural tube and generate different neural cell types. The anterior part of the neural tube develops into forebrain that adjoins the midbrain, followed by the hindbrain, and the posterior part of neural tube develops into the spinal cord (Figure 1A). In the dorso-ventral patterning, the neural tube is divided into a defined set of compartments, each containing distinct progenitor cells. Later on, each progenitor pool gives rise to a molecularly and functionally distinct class of interneurons or motor neurons. The patterning along the DV axis is initiated by activities provided by two signaling centers (Figure 1B). Ventrally, the notochord and floor-plate cells secrete the molecule sonic hedgehog (Shh). The concentration of Shh varies from ventral to dorsal axes. It induces the ventral subtypes and represses dorsal fates. Dorsally the roof plate secretes BMPs and Wnt which repress ventral identities and induce dorsal cell fates. At intermediate regions of the neural tube, retinoid signaling emanating from the somites adjacent to neural tube induces the generation of interneuron subtypes at this level [8].

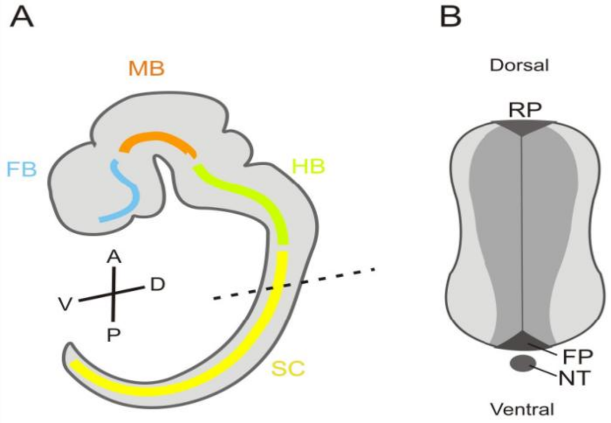


Figure 1A) Mouse embryo is subdivided along the AP axis: forebrain (FB), midbrain (MB), hindbrain (HB) and spinal cord (SC) and B) The neural tube is patterned along the DV axis: notochord (NT) and floor plate (FP) produce Shh; roof plate (RP) produces BMPs and Wnts [33]

In the ventral hindbrain, the high levels of Shh signaling establish a progenitor domain located dorsally to the floor plate, which expresses the homeodomain transcription factor Nkx2.2. During development, Nkx2.2 progenitors sequentially generate visceral motor neurons, serotonergic neurons and oligodendrocyte precursors in a time defined manner (Figure 2) [1,2]. During the period of VMN neurogenesis, Nkx2.2^{+} expresses the paired-like homeodomain transcription factor Phox2b. This is an important determinant of VMN fate [1]. At later stages, Nkx2.2$^+$ progenitors cease to generate VMNs and begin to produce 5HTNs. This is accompanied by the down-regulation of Phox2b and expression of high levels of Foxa2 [1,9]. Loss-of-function experiments revealed that loss of Phox2b results in the premature generation of 5HTNs [1]. In loss of Nkx2.2 mutants, 5HTN are not generated which is associated with an unexpected prolongation in the production of VMNs and a failure to suppress the progenitor expression of Phox2b and Nkx2.9 [1]. Shh also induces the expression of Tgfβ2 in Nkx2.2$^+$ progenitors at later stages. Tgfβ2 and Phox2b establish cross-repressive interactions which are important to establish a period of VMN generation at early stages and a robust repression of VMN production later [10]. This process can be recapitulated in ES cell bases *in-vitro* systems (Figure 2). Compared to *in-vivo, in-vitro* systems offers wide benefits to perform experiments for genome wide studies. Mouse ES cells are pluripotent cells obtained from the inner cell mass of a 3.5 day old embryo (the blastocyst). They can be maintained *in vitro* for extended periods without loss of their capacity to contribute to cell lineages when re-implemented back into blastocysts. ES cells have the ability to maintain a normal karyotype for innumerable cell divisions. The possibility to extract large amounts of mRNA from ES cells *in vitro* compared to *in vivo* makes ES cells an attractive model for genome wide analysis.

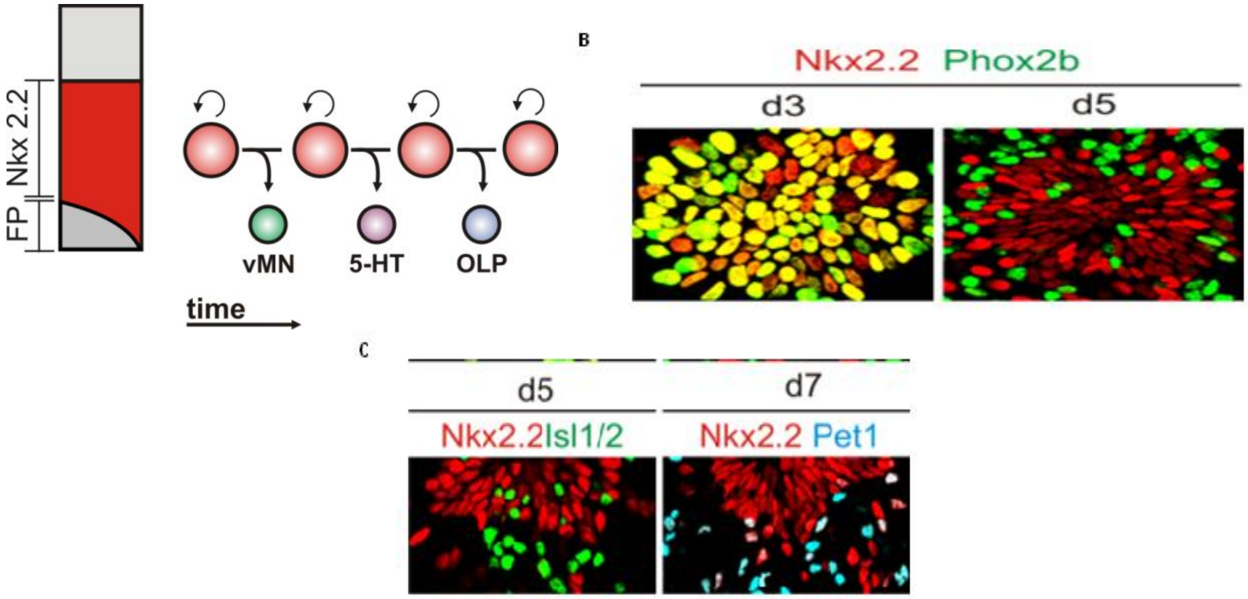


Figure 2 A)Sequential generation of visceral motor neurons, serotonergic neurons and oligodendrocyte precursors in temporal order B) Expression of Nkx2.2 and Phox2b at d3 (yellow), Down-regulation of Phox2b at d5 C) Generation of VMN (d5) from post-mitotic neurons detected by Isl 1/2 and generation of 5HTN at d7 detected by Pet1markers[10]

2.1 RNA-seq

High throughput sequencing technologies have recently become a popular methodology, commonly used among other applications to measure global expression with high accuracy. The technology of sequencing transcribed RNA followed by transcript mapping and quantification is known as RNA-seq. It has several solid advantages over micro-arrays and other previously developed methods. It is the first sequencing method that potentially allows the entire transcriptome to be surveyed in a very quantitative and high-throughput manner [11]. RNA-seq enables investigation of the complex aspects of transcriptomics, e.g. discovery of novel RNA entities, transcript isoforms, allele specific gene expression, verification of candidate mutations in RNA-coding regions, identification of DE genes between two conditions such as treated versus non-treated cells, cancer versus normal cells and between wild type and mutant strains.

The procedure of RNA-sequencing includes taking samples of purified RNA, shearing it, converting to cDNA, and sequencing on high throughput platforms such as Illumina Genome Analyzer, Applied Biosystems Solid or Roche 454 Life-sciences sequencing systems. This process can generate millions of short reads (25-700 bp) taken from one end of the cDNA fragments. Short reads could also be generated from both ends of each cDNA fragments (paired-end reads). After sequencing, the reads are mapped to the reference genome or transcriptome. The task at this step is to find the genomic location where each short read best matches to the reference genome, while allowing errors and structural variation. Next, the mapped reads are assembled into gene-level, exon level or transcript level summaries depending on the aim of the study. Later the summarized data are normalized for gene length and total transcript amount in the sample. Then the expression differences are compared with statistical tests, leading to a ranked list of genes with associated p-values, false discovery and fold changes estimates (discovery of differential expression). Further, biological insight from these lists can be gained by systems biology approaches [12].

2.2 Pathway/Enrichment Analysis

Comprehensive and insightful characterization of gene sets altered in specific conditions (AGS) is a challenging task. One of the most common approaches is to access the functional associations between a gene set of interest such as differentially expressed genes and known gene sets representing biological processes (e.g. GO terms) or pathways, generally termed as FGS, i.e. lists of genes that were previously assigned a common biological annotation. To identify and rank such associations, a wide range of enrichment analysis tools have been developed in recent years. The term enrichment analysis refers to examination of the list of genes to determine if they are over-represented among any set of certain processes or pathways members. Various enrichment analysis tools such as GSEA, DAVID, GoToolbox, and FATIGO etc address various challenges of functionally analyzing large gene lists. All these methods systematically evaluate the relationships between AGS and FGS, then statistically highlights the most enriched (over/under represented) biological annotations out of thousands of linked terms and contents.

Problem

The GO and KEGG databases do not encompass all functionally coherent groups and if the information about the differential expressed genes is not present in such databases, then it results in poor overlap between two sets and the analysis gives false negatives.

2.2.1 Network Enrichment Analysis

The idea behind NEA is to overcome the above mentioned limitations and consider the whole list of differentially expressed genes which are not necessarily to be the members of any already known functional category, but could be connected to such members in a network (Figure 3). Comparing to other tools, NEA [13], provides a network-based approach to illustrate novel gene sets with its biological functional categories. This method provides the interactions between AGS and FGS along with the comprehensive statistical evaluation. NEA is implemented in Perl (in preparation), R [13] and C++ [14]. The tool generates the statistics about the significant number of connections between two given gene sets.

The network enrichment was estimated with NEA z-scores. The standard z-score for the biological network connectivity between differential expressed gene list (A) and genes of a known functional group (F) was computed from the observed and expected link counts and their standard deviation as follows

$$\boldsymbol{z\text{-}score} = \frac{\boldsymbol{n_{AF} - \overline{n}_{AF}}}{\boldsymbol{\sigma_{AF}}}$$

In the given network, n_{AF} represents the total number of links between any genes of A and any genes of F, $\boldsymbol{\overline{n}_{AF}}$is the mean and σ_{AF} is the standard deviation. A default statistic counts the direct links between two gene sets. Under the true null hypothesis, i.e. in absence of any systematic functional links between gene groups, the z-scores should be approximately normally distributed. The z-scores could be converted to specific NEA p-values (the probability of a non-existing FGS-AGS relationship to be detected as existing by the NEA test) and to FDR (the probability of the detected FGS-AGS relationship to be genuinely false) by standard procedures. For both direct and indirect links, false discovery rates (FDR) were determined by adjustment for Benjamini and Hochberg multiple testing method [15]. Alternatively, the false discovery proportion was controlled by permutation tests on random gene sets of matching size and topological properties, which did not show any strong deviation from the Benjamini and Hochberg method.

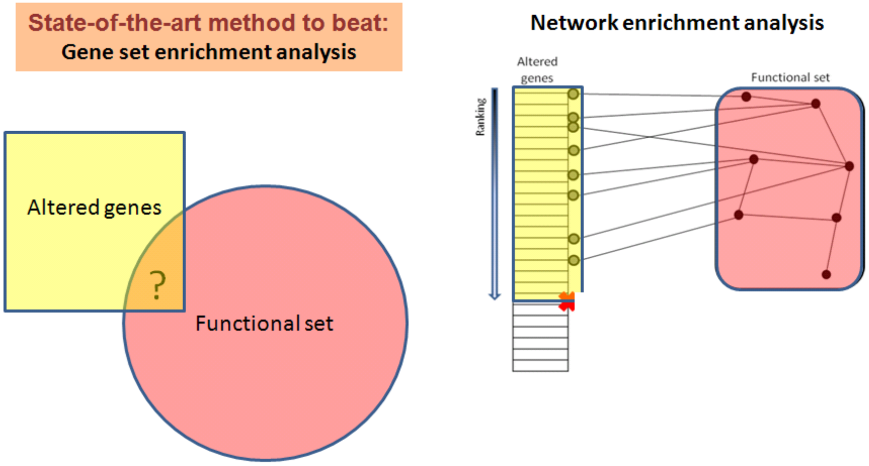

State-of-the-art method to beat:
Gene set enrichment analysis
Altered genes
?
Functional set
Network enrichment analysis
Altered genes
Ranking
Functional set

Figure 3: Figure illustrating the idea of GSEA and NEA [13]

For comparison with NEA, the GSEA scores were also calculated using the same AGS and FGS lists as in NEA with the hyper-geometric test, also known as Odds ratio test [16]. These z-scores were also converted to p-values and adjusted for multiple testing using Benjamini and Hochberg method.

The gene regulation and functional coupling (relation between gene/proteins in a network) in general is not limited to the transcription factor activity and there are various molecular mechanisms such as protein phosphrylation, mRNA coexpression, miRNA regulation etc. Funcoup [17] enables data integration from various experimental platforms which convey information about this variety of mechanisms to predict network links between nearly all genes and proteins in the global biological network. The latter thus includes differential expressed genes and pathways that determine different functional responses.

3 METHOD

The project is divided into two sections 1) Gene expression analysis of RNA-seq data 2) Enrichment Analysis. All the analysis was performed under the UNIX environment and programming languages used in this project were Perl, html and R-a statistical tool along with Bioconductor packages.

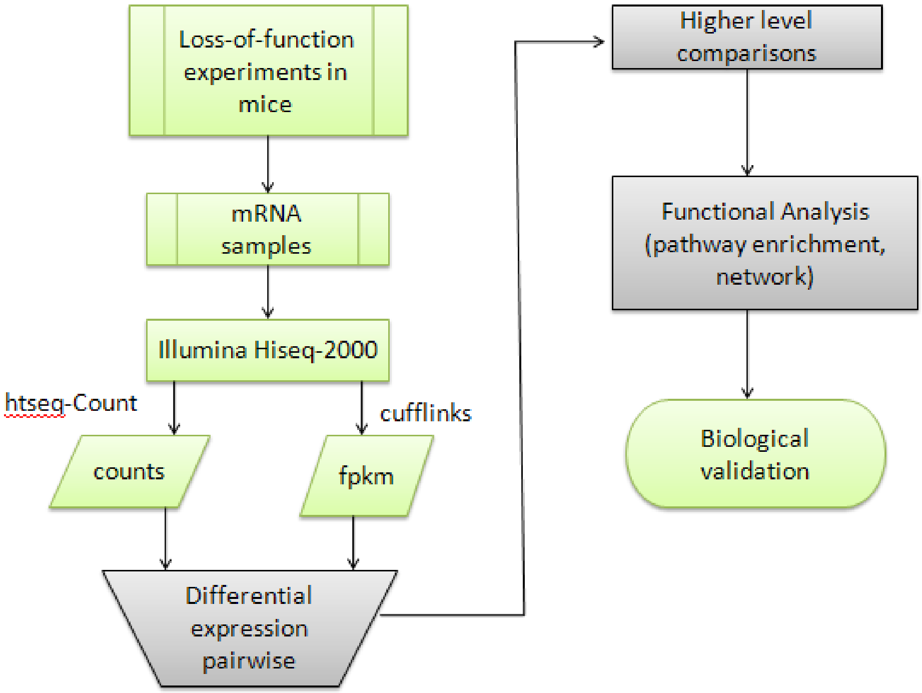

Loss-of-function experiments in mice
mRNA samples
Illumina Hiseq-2000
htseq-Count
cufflinks
counts
fpkm
Differential expression pairwise
Higher level comparisons
Functional Analysis (pathway enrichment, network)
Biological validation

Figure 4: The overall view of the data analysis. Boxes in grey are the ones described in this work.

3.1 RNA-Extraction

To determine the overall transcriptional changes associated with the temporal shifts in neural progenitor cells, loss-of-function experiments were done with mouse ES cells *in vitro* [Figure 5]. Analysis of Nkx2.2 mutants reveals genome wide transcriptional alterations of cells which fail to execute the temporal switch and thereby inappropriately continue to produce MNs at the expense of 5HTNs during late stages. Conversely, loss of Phox2b results in loss of VMNs and early generation of 5HTNs. In addition to Nkx2.2$^+$ and Phox2b transcription factors, the role of signaling molecule Tgfβ2 was also studied using an inhibitor of this signaling pathway, SB-505124 [18]. It is a small inhibitory molecule which binds to the kinase domain of Tgfβ2 receptor (Tgfβr1) thereby blocking Tgfβ2 signaling. Inhibition of this signaling pathway results in a prolongation of Phox2b expression and generation of VMNs at the expense of 5HTN.

Extraction of RNA samples for NGS experiments

mRNA samples were extracted by performing individual experiments on ES cells. First, in normal scenario (WT) Shh and RA were added to the culture medium which will induce the expression of Nkx2.2 progenitors around day 2.5-3.5. VMNs and 5-HTN were formed at day 3.5 and day 5.5 respectively. By applying magnetic cell sorting (MACS), Nkx2.2 progenitors were separated from the post-mitotic neurons and mRNA was extracted from the cells through standard methodologies. Similarly, mRNA samples were collected individually from the Nkx2.2 mutant, Phox2b mutant and SB-treated progenitor cells using MACS. These samples were subsequently subjected to deep sequencing. Culturing of ES cells and extraction of mRNA samples were done in the Department of Cell and Molecular Biology, Karolinska Institute.

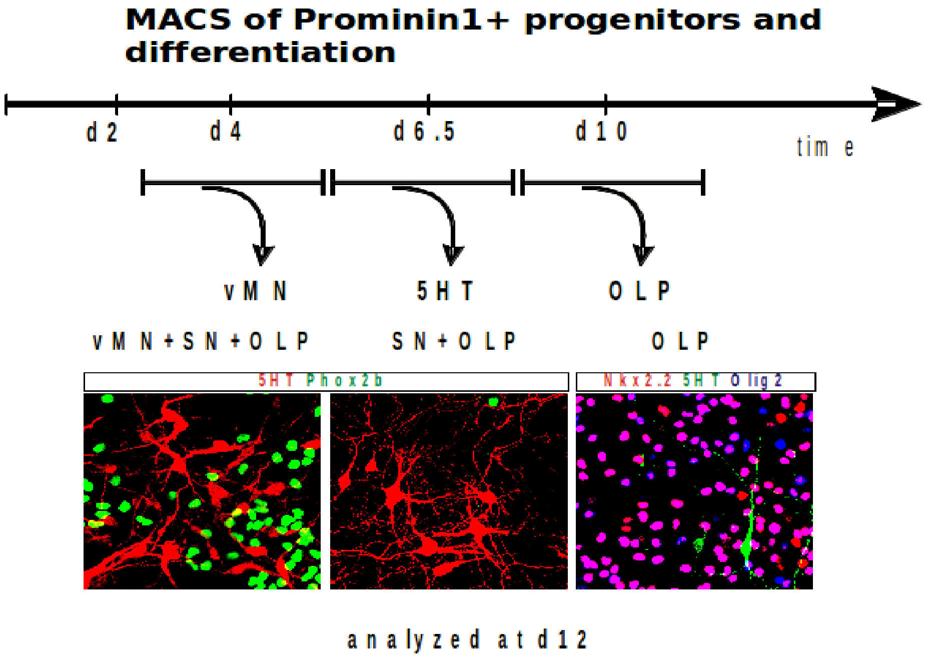
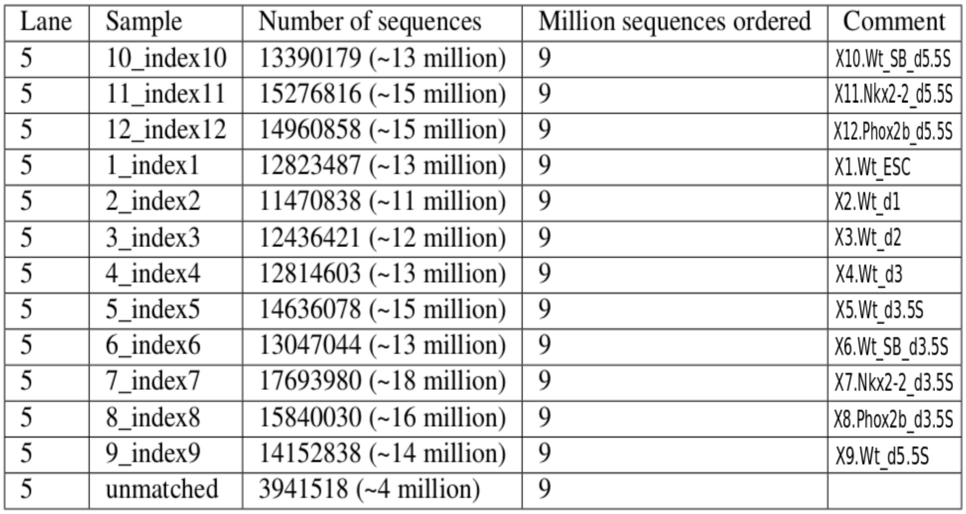


Figure 5: In vitro analysis of ESC differentiation. Differentiation of ES cells, appearance of VMNs around day 3.5, 5HTN (red) around day 5.5 and OLPs(pink) at day 9.5. Cells extracted from progenitor cells by MACS. [10]

3.2 RNA-seq Experiments

RNA–sequencing experiments were carried out on the second generation Illumina HiSeq-2000 machine at Science for Life Laboratory, Stockholm. The 12 samples extracted at different time points and experimental conditions were multiplexed and loaded into one lane of the sequencing machine [19]. The raw reads (~100 bases long) were analyzed with the tool Tophat to map them to the mouse reference genome as transcripts, taking into account exon-exon splice junctions [20]. This program takes the raw reads (fastQ files) as an input and produces BAM files. The BAM files (the aligned reads) can be viewed with the help of SAM tools. After the mapping procedure, a tool HT-Seq-Count [21] was used to obtain count values for genes, i.e. numbers of mapped raw reads. These gene-expression values were not normalized for the gene length, hence could not be interpreted directly as mRNA expression levels. E.g. longer transcripts have higher read counts at the same expression level than shorter ones. Furthermore, the samples differed in the total number of sequences (Table 1). Highest number of reads (around 18 million) was observed for Nkx2.2 mutant sample at day 3.5 whereas lowest number of reads (around 11 million) was noted for embryonic wild-type sample at day 1. In order to eliminate the gene length and sample biases, normalization was done by software Cufflinks [12,22]. The transcript abundances were represented as Fragment Per Kilobase of exon per Million fragments mapped (FPKM), i.e. values normalized by transcript length and total number of reads per sample.

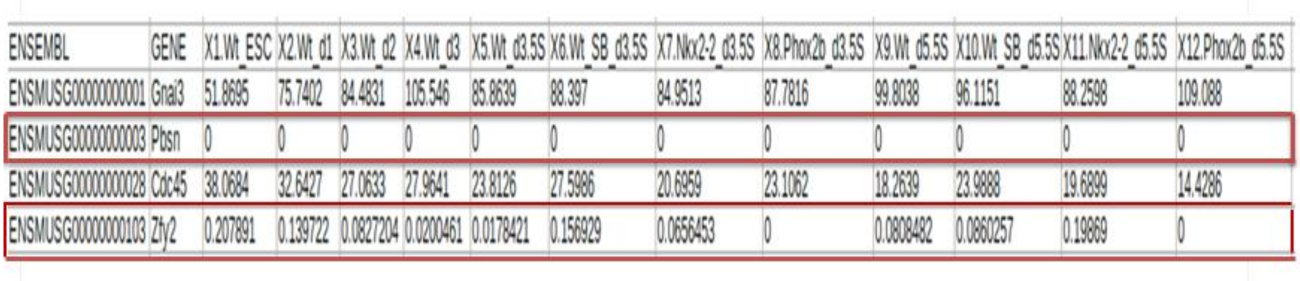


Lane	Sample	Number of sequences	Million sequences ordered	Comment
5	10_index10	13390179 (~13 million)	9	X10.Wt_SB_d5.5S
5	11_index11	15276816 (~15 million)	9	X11.Nkx2-2_d5.5S
5	12_index12	14960858 (~15 million)	9	X12.Phox2b_d5.5S
5	1_index1	12823487 (~13 million)	9	X1.Wt_ESC
5	2_index2	11470838 (~11 million)	9	X2.Wt_d1
5	3_index3	12436421 (~12 million)	9	X3.Wt_d2
5	4_index4	12814603 (~13 million)	9	X4.Wt_d3
5	5_index5	14636078 (~15 million)	9	X5.Wt_d3.5S
5	6_index6	13047044 (~13 million)	9	X6.Wt_SB_d3.5S
5	7_index7	17693980 (~18 million)	9	X7.Nkx2-2_d3.5S
5	8_index8	15840030 (~16 million)	9	X8.Phox2b_d3.5S
5	9_index9	14152838 (~14 million)	9	X9.Wt_d5.5S
5	unmatched	3941518 (~4 million)	9	

Table 1: Output of Illumina HiSeq-2000 machine. Column 1: lane number; column 2: sample index. Reads that had lost the sample indexing fragment (hence not matched) are given as unmatched. Column 3: number of sequences in each sample.

3.3 Gene-Expression Analysis

The first step of gene-expression analysis is to check statistics for all the samples. The data shown below is an example of fpkm table (normalized counts). The first column is Ensemble Id, second column is gene name and following columns represent the count value for all samples. While mapping, few genes were not properly assigned to the reference genome such gene values across the samples were given as 0 and some of the genes were expressed in low measure. Such genes were not considered for further analysis and were pruned from the list of DE genes.



ENSEMBL	GENE	X1.Wt_ESC	X2.Wt_d1	X3.Wt_d2	X4.Wt_d3	X5.Wt_d3.5S	X6.Wt_SB_d3.5S	X7.Nkx2-2_d3.5S	X8.Phox2b_d3.5S	X9.Wt_d5.5S	X10.Wt_SB_d5.5S	X11.Nkx2-2_d5.5S	X12.Phox2b_d5.5S
ENSMUSG00000000001	Gnai3	51.8695	75.7402	84.4831	105.546	85.8639	88.397	84.9513	87.7816	99.8038	96.1151	88.2598	109.088
ENSMUSG00000000003	Pbsn	0	0	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000000028	Cdc45	38.0684	32.6427	27.0633	27.9641	23.8126	27.5986	20.6959	23.1062	18.2639	23.9888	19.6899	14.4286
ENSMUSG00000000103	Zfy2	0.207891	0.139722	0.0827204	0.0200461	0.0178421	0.156929	0.0656453	0	0.0808482	0.0860257	0.19869	0

3.4 Clustering

The Clustering is a common statistical technique for exploratory data analysis. Clustering methods are divided into Hierarchical Clustering and Partitioning. The former is based on assumption of the hierarchical data i.e. genes/samples can be ranked based on their similarity whereas partitioning methods (like K-means) iteratively converges k-random initial clusters and assigning patterns to the nearest cluster [23]. Hierarchical clustering [24] was used for analysis. This method is based on calculating the distances between each data value. The distance measure used is called Pearson correlation, it captures similarity of the expression curves of two items (genes or samples). The similarity measure (r_{ij}) is given as equation 1

$$r_{ij} = \frac{\sum_c (e_{ic} - e_i)(e_{jc} - e_j)}{\sqrt{\sum_c (e_{ic} - e_i)^2 \sum_c \left(e_{jc} - e_j\right)^2}} \text{--} (1)$$

The similarity score (r_{ij}) closer to 1 indicates a perfect correlation. Hierarchical clustering includes several methods, namely single linkage, complete-linkage, average-linkage, wards etc. The distinction among them depends on how the distances between clusters are calculated. Average method was used to calculate the distances between clusters; this method uses the average distance between objects from the first cluster and objects from the second cluster.

3.5 Differential Expression Detection

Although RNA-seq experiments are claimed to generate highly reproducible results (hence no replicates in many RNA-seq projects) it is difficult to eliminate the non-biological experimental variability from the true biological differences. To better infer true expression values, appropriate modeling of the variability is important. The formal aim of statistical testing is to try rejecting the so called null hypothesis, i.e. an assumption of no difference between the expression levels of two samples. The differential expression of a particular gene is reported as significant if the observed difference in expression levels is greater than what is expected just due to natural random variation with a certain probability.

There are numerous methods to measure gene expression from micro-array experiments. All those methods could not be applied directly to RNA-seq experiments because it gives count measurements (discrete nature), whereas micro-array experiments gives continuous intensity values. In micro-array experiments, the fluorescent intensity values are log transformed and then analyzed as normally distributed random variables, while transformation of count data is not well approximated by continuous distribution especially when dealing with lower count range or with small samples [12]. Hence, statistical methods such as t-test and linear models cannot be applied directly to analyze the micro-array data; indeed such methods are most suitable when dealing with replicates. In this study, genes of our particular interest were transcription factors that often have genuinely low expression levels and hence low counts which make differential expression less sensitive.

DESeq [25] is one of the most commonly used packages to do differential expression analysis for sequence count data. It assumes a non-linear relationship between the variance and mean expression levels which allows the variance to be fitted using pooled data with similar expression levels. This is done since low number of replicates make the estimation of variance difficult when using just the data available for a particular gene. It produces test p-values based on approximation of exact test (similar to fishers-exact test with changed calculations) [25]. By using the DESeq package only very few DE genes were discovered. Barplots, MA plots describing the p-value distribution obtained by DESeq is discussed in the Results Section (Figure 9A-C). It is important to note that the samples analyzed were without biological replicates. Hence, for this project we used binomial z-scores that employ odds ratios of observed fold changes to calculate the p-values and false discovery rates (Figure 9D-F). This scenario was different from the general cases, where the p-values can be calculated by applying linear models or t-test/ANOVA test.

3.5.1 Odd's (binomial) z-score

Odd's method yields very good results in addressing various biological questions. The odds is defined as ratio of probability that the event of interest occurs to the probability that it does not occur. It provides an estimate with confidence interval for relationship between the two binary (Absent/present) variables. Generally, Odds Ratio (Equation 2) is explained by 2x2 tables [16]. The zero counts of genes in all samples were removed by adding a pseudocount ‘0.5’ which reduces the number of tests to perform and false discovery rates.

	Experimental	**Control**
Gene i	**a**	**b**
Remaining genes	**c**	**d**

$$\mathbf{L = \log(\boldsymbol{ad/bc})} \quad \text{---------------} \ \mathbf{(2)}$$

L represents samples log odds ratio, a, b represent gene expression value for particular gene i and c, d represent summation of gene expression values for the remaining genes in mutant and WT samples respectively.

The distribution of log odds ratio is approximately normally distributed.

$$\boldsymbol{X \sim N\,(log\ (OR),\ \sigma^2)}$$

In order to test whether these odd ratios calculated were significant, standard error of log odds ratio were calculated as follows

$$\boldsymbol{SE} = \sqrt{1/a+1/b+1/c+1/d}$$

The z-scores (standardized log-odds ratio) = $\frac{\log Oddsratio\ (L)}{standard\ error\ (SE)}$

Using R function for two-sided test, p-values were calculated from z-scores by the formula, 2*(1-pnorm (abs (z-score)))

p-value

The p-value is defined as an estimated probability of rejecting null hypothesis (Ho) of a study question when that hypothesis is true. In our context, null hypothesis was no differences between the wild-type and mutant samples for a particular gene i. If the p-value is less than significance level (<0.05), then we rejected the null hypothesis i.e. accepted a difference exists between two samples for a particular gene i. In order to avoid low counts, the genes with count sum for a pair (si, sj) greater than 5 were considered to calculate the p-values from their binomial z-scores.

Multiple Testing Problem (False Discovery Rate)

When we set a p-value threshold of 0.05, it means there is 5% chance that the result is a false positive. While 5% was acceptable for one test; if we had done multiple tests, then this 5% would result in a large number of false positives which is called multiple testing problem. We applied Benjamini-Hochberg method [15] to control FDR and accepted the cutoff 0.05 of this adjusted p-value (q-value), which implies that 5% of DE cases might be false positives in reality.

Fold change

Fold change estimates represents the biological significance and is often used in gene expression analysis of micro-array, RT-PCR and RNA-seq experiments to measure the change in mRNA/gene expression level. It is a number in logarithm based 2 showing how many folds the gene expression has gained from wild-type(x) to mutant condition(y).

$$\text{Fold-change}_i = \log_2 {}^{y}\!/_{x}$$

Log2-Fold-changes	Up-regulated	Down-regulated
1.5 fold	+1	-1
2-fold	+0.585	-0.585

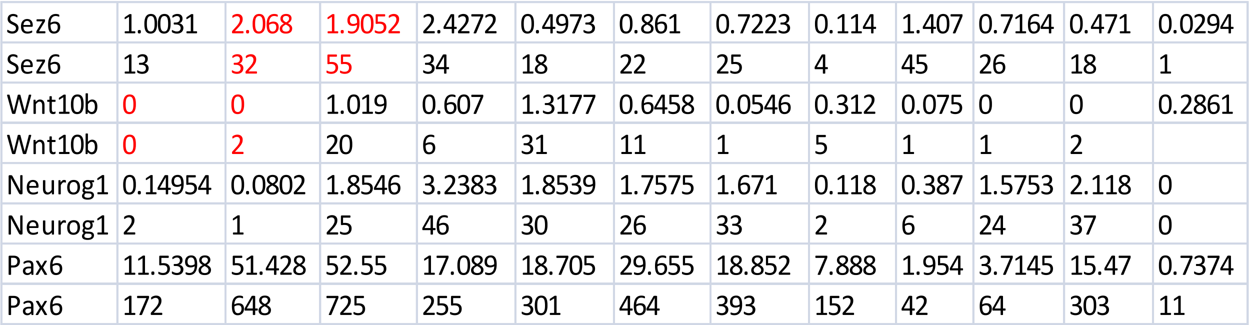
Log2-fold change values

The genes with fpkm sum and count sum for a pair (si,sj) greater than 5 (to avoid very low counts) were further considered for the fold-change estimation. In some situations for a gene, fpkm fold-change was high (high gene expression) and corresponding q-value was also high indicating lower statistical significance. E.g. gene Wdr17 has high fpkm log2-foldchange (1.938) between SB treated at day 3.5(x6) and wild-type at day 3.5(x5), but the q-value is 0.69. The count values of x5 and x6 were 11 and 14, indicating almost same gene expression. So, the binomial z-score will be close to 0, hence q-value is high. There exist many such cases in the data which leads to inconsistent results. The genes whose normalized and raw count fold-change values greater than threshold were considered for further analysis. The reason to consider both the normalized and raw counts is to make the analysis too strict.

Correlation

To find the degree of similarity between fpkm and count values, correlation analysis was done. Correlation values between fpkm and count values for each gene were calculated with an R function using Pearson method (Equation 4). Figure 6 represents the correlation histogram, where each bin corresponds to total number of the genes and x-axis represents the degree of correlation. In correlation analysis +1 indicates strong correlation, whereas -1 indicates negative relation between the fpkm and count variables. The bins below 0.0 were genes whose fpkm and count values were not related, Table 2 gives the information of such unrelated genes. E.g: gene Sez6, for sample 2 the values of fpkm and counts were 2.068 and 32 respectively, whereas for sample 3 values were 1.9052 and 55 respectively. This shows the disagreement between two variables. In reality count value of sample 2 should be greater than samples 3 since the fpkm value of sample 2 is greater than sample 3. This variability is explained as read mapping problem in RNA-seq experiments where mapping was done by two different programs which have different policies. The correlation factor was taken in account to eliminate count-fpkm bias and only the genes with correlation value greater than 0.5 were considered for further analysis.

$$\mathrm{r} = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x^2)][n\sum y^2 - (\sum y^2)]}} \quad \text{---------------} \; (4)$$



Sez6	1.0031	2.068	1.9052	2.4272	0.4973	0.861	0.7223	0.114	1.407	0.7164	0.471	0.0294
Sez6	13	32	55	34	18	22	25	4	45	26	18	1
Wnt10b	0	0	1.019	0.607	1.3177	0.6458	0.0546	0.312	0.075	0	0	0.2861
Wnt10b	0	2	20	6	31	11	1	5	1	1	2	
Neurog1	0.14954	0.0802	1.8546	3.2383	1.8539	1.7575	1.671	0.118	0.387	1.5753	2.118	0
Neurog1	2	1	25	46	30	26	33	2	6	24	37	0
Pax6	11.5398	51.428	52.55	17.089	18.705	29.655	18.852	7.888	1.954	3.7145	15.47	0.7374
Pax6	172	648	725	255	301	464	393	152	42	64	303	11

Table 2: Values in red represent uncorrelated fpkm and Counts. 2-13 columns represent the various samples at particular condition. Fractional values represent fpkm values whereas numeric values represent count values.

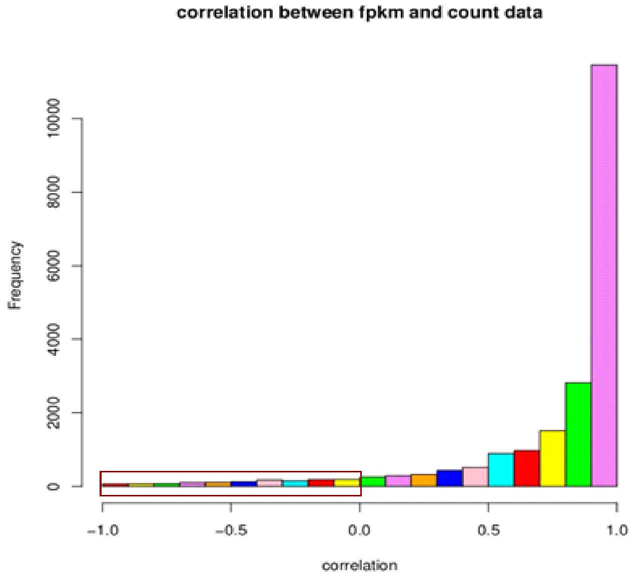


Figure 6: Correlation between fpkm and count data. Height (y-axis) of bar represents number of genes, x-axis represents degree of correlation

Fpkm Sum

Some of the genes across samples had very low fpkm values i.e. 0 to 10. It might be due to a sequencing artifact where reads were not properly mapped to the genes. Instead of eliminating all the genes with low fpkm values, genes that had sum of fpkm values greater than cut-off and genes with good correlation between fpkm-counts were considered. From (figure 7 c), the cut-off value was defined as 10 since bins showed a positive correlation (greater than 0)

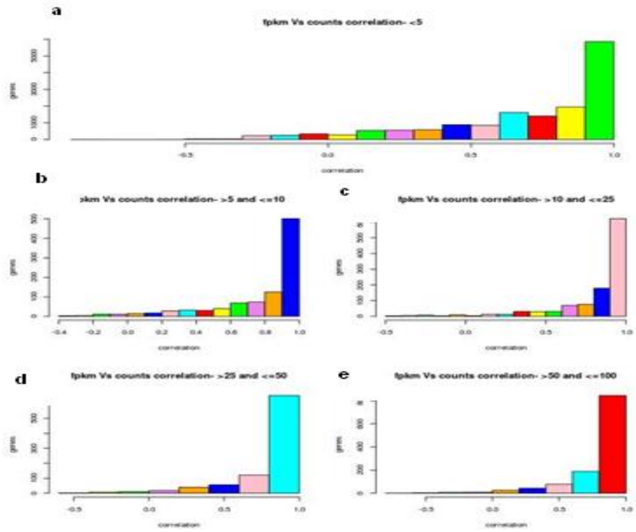


Figure 7: Histograms plotted against fpkm_sum and fpkm-count correlation

3.6 Cross-comparisons between WT and mutant samples (Quadruplet Comparisons)

The Venn-diagrams give a clear representation of data, like how many genes were differentially expressed and if there were any overlapping genes between pairs of samples (e.g. x9x5, x8x5, etc). 'VennDiagram' package in R was used for plotting; it is a flexible package for making two to five comparisons. In order to identify the differentially expressed genes across samples, quadruplet comparisons were done. The DE genes which passed the above mentioned criteria were selected and Venn-diagrams were generated. Venn-diagrams just give numerical representation of data, but it does not give information of which genes were expressed. Further in-order to access the gene information in particular comparison, an R code was written to generate two separate text files for 1.5 fold difference and two-fold difference, which can be easily accessed to visualize the genes of interest based on their fold-change values. The files contain information about the Ensemble Id, gene names and its description, normalized expression values, correlation value, p-adjusted values, fold-change values, so on. Besides these, there is column with condition. For example in a quadruplet comparison between four cases (x6x5,x9x10,x11x8,x8x5), the genes which pass the criteria in a particular case was assigned as “+” , those genes which did not pass was set to “.” . If a gene is passed in all four cases, the condition was given as “++++”. So with this condition as key, it is possible to extract the genes and its information for a particular case from the text file.

3.6.1 High-level cross comparisons between WT and mutant samples

From the above analysis, an approach to represent the DE genes was identified. Applying the same protocol, high-level cross comparisons were done to identify novel genes involved in VMN-to-5HTN fate switch.

The block of TGFβ2 signaling by SB inhibitor as well as knockout of Nkx2.2 should extend the period of VMN generation in the WT neural progenitors and delay the 5HTN generation. By the selection of genes ‘up-regulated’ during differentiation in the control conditions and the same time not changing the expression level in the SB treated samples or Nkx2.2 knockout cells , the important regulators of 5HTN generation can be targeted. Conversely, the genes showing ‘down-regulation’ during the differentiation but maintaining high expression level upon TGFβ2 inhibitor treatment or in Nkx2.2-deficient cells assumed to be important for the generation of VMN (Figure 8A). Phox2b mutant progenitors are not able to produce VMN at any stage of differentiation. In the mutant cells at the time, when neurogenesis is initiated, 5HTN are generated instead of VMN. Thereby the genes ‘upregulated' in the Phox2b mutant progenitors in comparison with the WT progenitors at d3.5 could be involved in 5HTN generation. The same time these genes should show significant up-regulation in the WT cells from d3.5 to d5.5 of differentiation – with the changes in the WT progenitor cell competence from VMN to 5HTN generation. Comparing the opposing list of genes down-regulated in the WT cells during the differentiation and the same time showing low expression level in the Phox2b mutant progenitors in comparison with WT cells at d3.5, the aim is to find genes important for the VMN fate (Figure 8B).

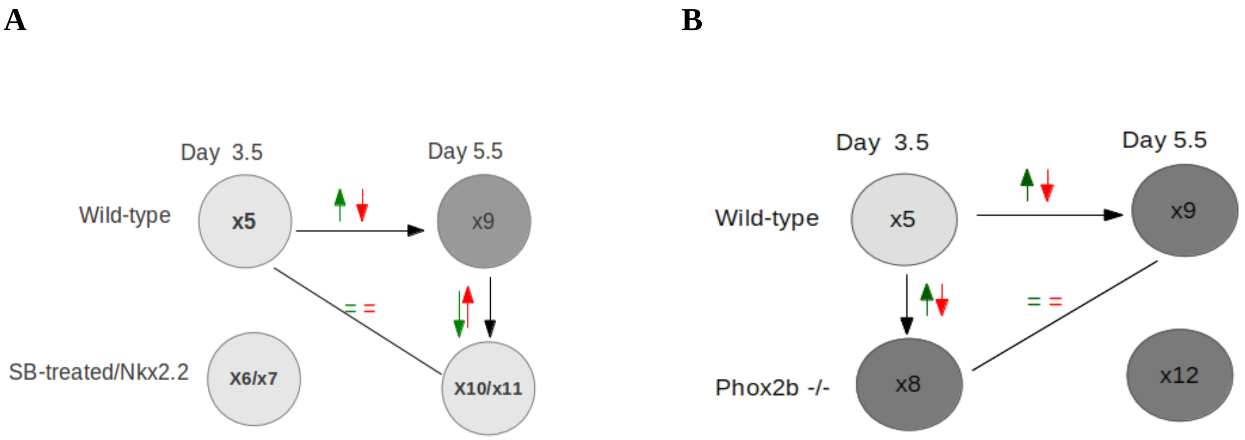


Figure 8: Formation of neurons on defined time points. Circles with light grey indicate 'VMNs' and circles with dark grey indicate '5-HTN'. Arrow represents pattern of regulation (up/down).A) Comparison of wild-type with SB-treated/Nkx2.2 samples and B) Comparison of wild-type samples with Phox2b samples.

3.7 Enrichment Analysis

3.7.1 Gene Set Enrichment Analysis

GSEA [26] is one of the popular methodologies to identify previously known functional gene sets (GO terms) from the list of differential expression genes that are over-represented either in top or bottom of the ranked list based upon enrichment scores. Enrichment scores were given by walking down the ranked list of genes, increasing a running sum when a gene is in the gene-set and decreasing when it is not present. A positive ES score indicates gene-set enrichment at the top of the ranked list and negative

ES score indicates gene set enrichment at the bottom of the ranked list. In the study, GSEA was performed using Pre-ranked mode, where the list of genes was ranked according to the fpkm log2 fold change values. Rather than p/q-values, fold-change values were considered to detect the enrichment peak.

3.7.2 Network Analysis

The idea of NEA is to find relationship between the AGS and FGS by providing the number of links between two gene sets.

AGS: List of differential expressed genes, whose function has to be identified

FGS: List of known biological pathways, 8024 groups from KEGG pathways [27] and Gene Ontology terms (biological process domain) [28] were considered.

Network: Funcoup's mouse network was considered as default network. For the analysis, it was merged with KEGG pathway links, members of protein complexes (CORUM database) and a smaller network of gene regulation which was reverse-engineered from a data set of transcription factor knock-down [29]. In total, it contained 1007449 links between 18141 genes. The network tool generated links with default parameters - 3 iterations, no cut-off value. The output is number of connections between the two gene sets and respective statistical confidence p-values and FDR. The output was given as an input text network file for Cytoscape [30], setting both network edges and edge attributes. Cytoscape is network visualization and analysis software which provides a basic functionality to layout and query network. In Cytoscape genes are represented as nodes and interactions between two genes were represented as links (edges)

4 RESULTS

Initially 37682 genes (raw counts) were identified as differentially expressed from the RNA-seq experiments, but after data normalization (fpkm) and removing duplicates, it was narrowed to 37514 genes. All this data were stored in TAB- delimited text file (.TXT) and can be accessed via Microsoft Excel or any other spreadsheet application. Figure 9, shows how the samples were clustered well based on their time points i.e. X1...X4 belongs to one cluster and samples from day 3.5 (X5...X8). Among samples from day 5.5, X12 looks different which indicates that Phox2b sample on day 5.5 has different expression level compared to the other samples on day 5.5.

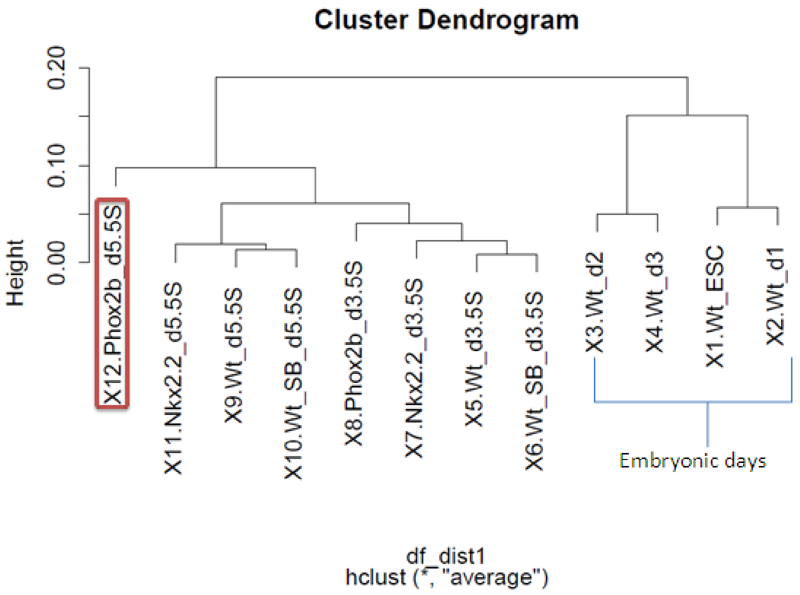


Figure 9: Hierarchical sample clustering

4.1 Comparison of Binomial z-score method and DESeq

The differential expression attributes such as p-values, fdr and fold change values from fpkm and count data were calculated using DESeq package, implemented within R Bioconductor. The first plot (Figure 10A), describes the variance (dispersion) distribution. If the means across the genes were equally distributed, the dispersion trend (red line) fits the dispersion dots and will be a straight line, but in the dataset there exists dispersion for the genes with counts above 100. The second plot (Figure 10B) displays the p-value distribution. In theory given an adequate statistics and complete absence of differential expression, a p-value distribution should be uniform, i.e. have bins of nearly equal height from 0 to 1. Given some differential expression, it should have a peak in the region close to 0, and rest of the distribution should be flat. However, in the DESeq case, it showed an opposite behavior which means that the statistical test was not correct, or a wrong variance model was used.

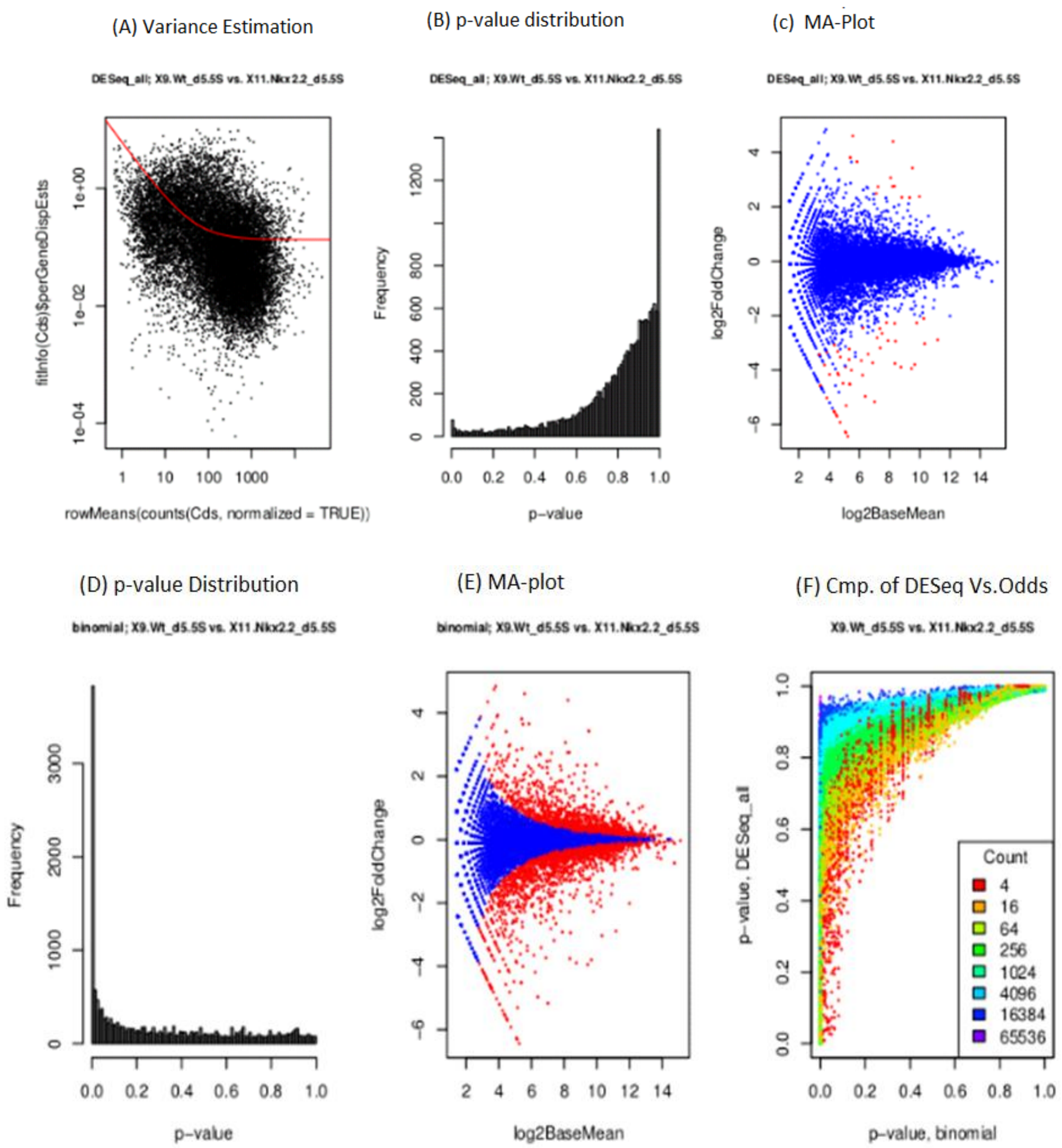


Figure 10: A-C: DE analysis by DESeq package; D-F: DE analysis by Odds method

The third plot (Figure 10c) is the scatter plot of log2 ratio (fold change) versus means. The red dots represent genes detected as differentially expressed, so only very few differentially expressed genes were detected for the dataset. In other words, this method was not appropriate for this data; instead the binomial z-test (Odds method) was used. Figure 10D, represents the p-value distribution plot and MA plot (Figure 10E) where p-values were calculated by Odds method as described in the methodology. The distribution follows real distribution order, where more significant p-values were found at 0. A comparative analysis of DESeq and Odds methods was made and generated p-values with respect to counts (Figure 10F). It clearly shows that genes with lower counts values were detected by both the methods, whereas higher count genes (count=16384...65536) were not detected by DESeq package.

4.2 Selection of Differential Expressed genes

The statistically and biologically significant genes were selected by considering all the factors as mentioned in METHODS. In the analysis, a particular gene ‘i' is defined as differentially expressed (+) either up-regulated or down-regulated, if it satisfied the following set of criteria

1. Correlation greater than 0.5, to overcome fpkm-count correlation bias.
2. Pairwise sum of FPKM values (si, sj) greater than 10, because fpkm values having 0 and low values indicates low expression, hence were of less interest.
3. FDR (p-adjusted) <=0.05; to avoid false positive estimation.
4. log2-foldchange (fpkm) > 1.5 or 2-fold and log2-foldchange (counts) > 1.5 or 2-fold to take into account biological significance

Based on the set of criteria, quadruplet comparisons across WT and mutant samples at various time points were done (Appendix C). Venn-representation of the number of genes identified across the WT and SB treated samples is shown in figure 11.

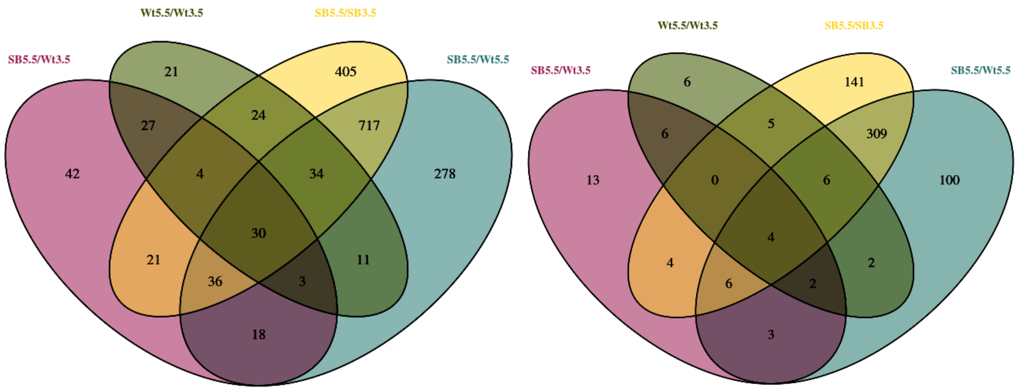


Figure 11: Venn representing quadruplet comparisons a) 1.5 fold b) 2-fold

4.2.1 Web-Page

In order to make the analysis more clear and accessible to biologists, an interactive web page was developed based on html and java functions. The webpage [31] contains information about quadruplet comparisons of differentially expressed genes represented as interactive Venn-diagrams. When user clicks on the number, it gives the genes that are expressed in that context. The information about a gene can be viewed by clicking on it which further redirects to Mouse Genome Informatics (MGI). Moreover, gene expression can be known from the small color boxes (red: down-regulated, green: up-regulated) adjacent to the gene name which corresponds to fpkm log2foldchange values.

4.3 Genes associated with VMN and 5HTN generation

High-level cross comparisons as described in METHODS chapter was analyzed on the WT and mutant samples and potential DE genes associated with the VMN/5HTN were narrowed down with statistical and biological significance. From the list of identified genes, those genes which were associated with the regulatory functions such as DNA binding, transcription events were highlighted.

4.3.1 Comparisons between WT and Nkx2.2 samples

97 genes were found to be associated with 5HTN generation by cross-comparative analysis between the genes that were upregulated during differentiation in WT cells over time and downregulation in Nkx2.2 mutant cells. Contrary, by cross-comparative analysis between genes those were downregulated in WT progenitors and upregulated in Nkx2.2 mutant cells to WT cells at d5.5 identified 43 genes that were related to VMN generation. Among the 43 genes, two genes Nkx2-9 and Nkx6-2 were linked to the potential regulatory terms, and from 97 genes stat2, Gfra2, Npy2r, and Lrn1 were found to be associated with the regulatory activities.

Samples	Up-regulated		Down-regulated	
	1.5 fold	2-fold	1.5 fold	2-fold
Wild-type (day3.5->day 5.5)	**805**	**330**	**466**	**84**
Nkx2.2-/- (day5.5->wild-type 5.5)	**245**	**145**	**234**	**102**

Table 3: Over-expressed genes identified by wild-type and Nkx2.2 -/- comparisons

4.3.2 Comparison between WT and SB treated samples

Cross-comparisons between the WT and SB-treated samples identified 58 common DE genes that were associated with the 5HTN generation and 7 common DE genes related to VMN generation. Very few genes were identified between downregulated genes during differentiation in WT and upregulated genes in Tgfβ2 inhibitor treatment samples. Among the 58 genes that were found to be associated with 5HTN generation, 4 genes (Table 4B) show connection with the potential regulatory functions.

(A)

Sample	**Up-regulated**		**Down-regulated**	
	1.5 fold	**2-fold**	**1.5 fold**	**2-fold**
Wild-type (day3.5 to 5.5)	**805**	**330**	**466**	**145**
SB-treated (day 5.5/ wildtype 5.5)	**27**	**1**	**127**	**30**

(B)

Eno2	**Enolase 2,gamma neuronal**
Npas3	**Neuronal PAS domain protein 3**
Nbl1	**Neuroblastoma, suppression of tumorigenicity 1**
Hoxb3	**Homeo box B3**

Table 4: A) Overview of total number of DE genes expressed by cross-comparisons between wildtype and SB treated samples. Colors indicate one set of analysis for pairwise comparisons B) Potential regulatory genes identified from 58 common DE genes

4.3.3 Comparison between WT and Phox2b samples

The cross-comparisons done between the upregulated genes in Phox2b mutants with the WT progenitors at d3.5 and genes that were significantly upregulated in the WT cells over time gave 102 DE genes (1.5 fold). Conversely the analysis done on opposite list of genes showed 53 DE genes (1.5 fold). 102 and 53 identified genes were associated with the generation of 5HTN and VMN respectively.

Samples	Up-regulated		Down-regulated	
	1.5 fold	2-fold	1.5 fold	2-fold
Wild-type (day3.5->day 5.5)	**805**	**330**	**466**	**145**
Phox2b-/- (day3.5->wild-type 3.5)	**236**	**79**	**305**	**101**

Table 5: Total number of over-expressed genes identified by wild-type type and Phox2b -/- comparison

A

Eno2	**enolase 2, gamma neuronal**
Slc6a15	**solute carrier family 6 (neurotransmitter transporter), member 15**
Nrcam	**neuron-glia-CAM-related cell adhesion molecule**
Optn	**Optineurin**
Foxa2	**forkhead box A2**
Nos1ap	**nitric oxide synthase 1 (neuronal) adaptor protein**
Ncald	**neurocalcin delta**
Hist3h2ba	**histone cluster 3, H2ba**

B

Olig3	**oligodendrocyte transcription factor 3**
Six1	**sine oculis-related homeobox 1 homolog (Drosophila)**
Barx2	**BarH-like homeobox 2**

Table 6: List of potential regulatory genes identified among A) 102 DE genes associated with B) 53 DE genes

4.4 Pathway Analysis

Different approaches of GSEA and NEA were tested to find an appropriate and suitable method among them to analyze the data.

4.4.1 GSEA Pre-ranked tool

The 203 genes that were up-regulated between Phox2b-/- at d3.5 and WT at d3.5 were tested using GSEA Pre-ranked tool. The analysis gave two enrichment lists based on their enrichment scores. Out of 83 gene-sets, 45 and 38 gene sets show positive and negative enrichment scores respectively. The ranked list obtained does not give proper statistical significant p-values (see appendix B).

Drawback: This tool is not suitable for sophisticated (more than pairwise) comparisons, because only one list of fold change values can be given for ranking.

4.4.2 Network Enrichment Analysis

To illustrate the biological significance from the pairwise analysis, NEA was applied to 257 common DE genes that are found to be associated with 5-HTN generation. The KEGG terms were filtered from NEA output and significant ones were extracted using the chi-square fdr< 0.05 and visualized in cytoscape [figure12]. When we look at the signaling pathways, mapk, wnt , calcium and gnrh signaling pathways were found to be associated with all three comparisons (red circles). Similarly, 103 common DE genes that are responsible for the formation of VMN identity were subjected to NEA (appendix D). Tgfβ2 signaling pathway was associated with the genes that were down-regulated in wild-type (x9x5), Phox2b mutant (x8x5) samples and down-regulated genes in wild-type (x9x5), up-regulated genes in Nkx2.2 mutant samples (x11x9). This analysis helps in the identification of signaling pathways that were regulated by the overrepresented genes in particular comparison or in mixed comparisons.

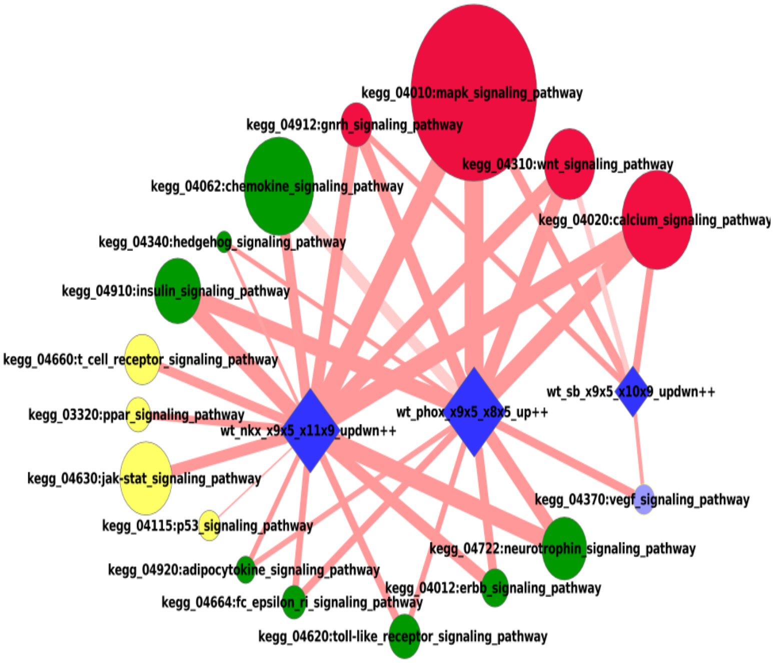
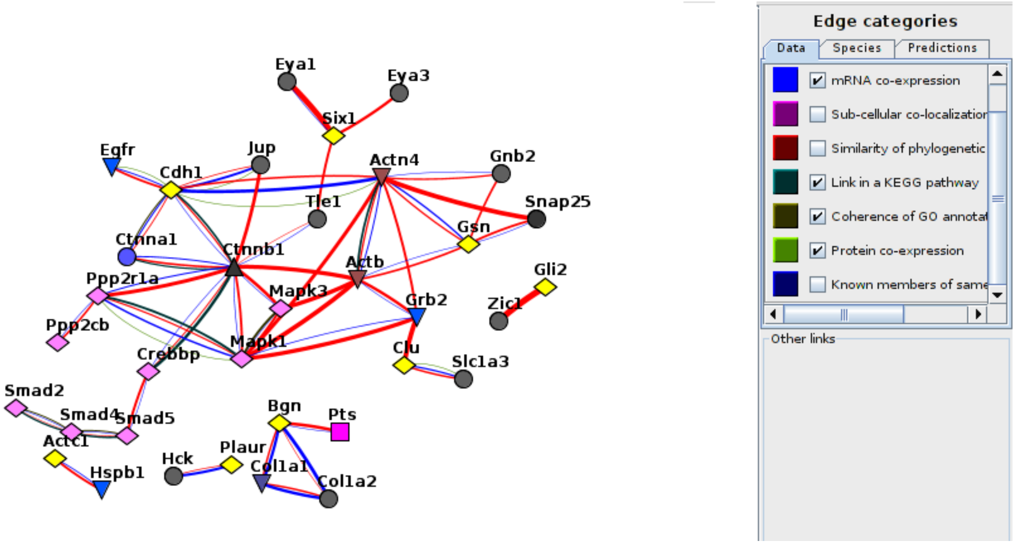

kegg_04010:mapk_signaling_pathway
kegg_04912:gnrh_signaling_pathway
kegg_04310:wnt_signaling_pathway
kegg_04062:chemokine_signaling_pathway
kegg_04020:calcium_signaling_pathway
kegg_04340:hedgehog_signaling_pathway
kegg_04910:insulin_signaling_pathway
kegg_04660:t_cell_receptor_signaling_pathway
wt_sb_x9x5_x10x9_updwn++
kegg_03320:ppar_signaling_pathway
wt_phox_x9x5_x8x5_up++
wt_nkx_x9x5_x11x9_updwn++
kegg_04630:jak-stat_signaling_pathway
kegg_04370:vegf_signaling_pathway
kegg_04115:p53_signaling_pathway
kegg_04722:neurotrophin_signaling_pathway
kegg_04920:adipocytokine_signaling_pathway
kegg_04012:erbb_signaling_pathway
kegg_04664:fc_epsilon_ri_signaling_pathway
kegg_04620:toll-like_receptor_signaling_pathway

Figure 12: Comparison of genes associated with 5HTN fate. The diamonds represents number of genes in ags, circular nodes represents number of genes in fgs, edge (links) color represents the chi-square fdr, edge line with represents the number of links between ags and fgs. Circles in red show enrichment with all three comparisons, circles in green show enrichment with wild-type VS Nkx2.2 and wild-type Vs Phox2b comparison. Circles in yellow show enrichment with WT Vs Nkx2.2

4.4.3 Funcoup Analysis

The genes in the system are also influenced by other interactions such as mRNA co-expression, protein-protein interactions which are involved the fate switch from VMN-to-5HTN. To find such links funcoup analysis was done on 13 TGFβ2 signaling related genes identified by NEA. The result (Figure 13) shows that all the 13 genes related to TGFβ2 (yellow nodes) was not interacting with each other, though the signal is collected at TGFβ2. The analysis helps to identify the nearby genes that were related to differential expressed gene.



Edge categories

Data | Species | Predictions

🟦 ☑ mRNA co-expression
🟪 □ Sub-cellular co-localizatior
🟥 □ Similarity of phylogenetic
▨ ☑ Link in a KEGG pathway
▨ ☑ Coherence of GO annotat
🟩 ☑ Protein co-expression
▨ □ Known members of same

Other links

Figure 13: Yellow, pink diamonds, AGS, FGS genes related to TGFβ2 signaling pathway. Blue lines represent mRNA co-expression; Redlines represent protein-protein interactions; Green line represent protein co-expression.

5 DISCUSSION AND CONCLUSION

With the advancement of high-throughput sequencing methods, gene-expression analysis by micro-arrays are being replaced by RNA-seq technology. In this project 12 samples were used to determine expression changes between wild-type cells and cells knocked out for specific transcription factors. (mutant samples). RNA-seq experiments generate huge amounts of data and statistical analysis is required to reduce false positive estimations and get desired results. Microarray's, pioneer in the gene expression analysis has many packages and well-defined methods to deal with differential expression data. However, such methods cannot be directly applied to RNA-seq. Methods based on negative binomial distribution (DESeq, EdgeR) developed to deal with RNA-seq might not be applicable to all datasets. Indeed each method performs better under different circumstances and the choice for method needs to be done according to the data itself. The data did not contain any biological replicates due to high cost so performing linear modeling is not possible. We implemented a different approach where we calculated p-values from the binomial z-score (Odd method). This method identified significant differential attributes from low counts and no biological replicates situation compared to DESeq. For very small counts, the normality and variance used by the Odds method might not be efficient. So, in the analysis p-values were calculated from the raw counts greater than 5 to avoid very low counts. Further, we would like to perform the analysis with fisher's exact test and compare with the results of odds method. After data normalization 37514 genes were identified as differentially expressed. To filter the differentially expressed genes based on statistical and biological significance besides q-values and fold-change cutoff's we considered additional sets of criteria to overcome the read-mapping issues and avoid very low expressed genes. The Venn-diagrams and tab-delimited text files give clear view on the differentially expressed genes and its related information. Further webpage that was developed based on the Venn-diagrams gives easy access to look at the differential expressed genes with respect to their fold-change expression values. High level cross comparisons performed across different combinations of wild-types and mutants helps in the investigation of new genes that are responsible for the VMN-to-5HTN cell-fate switch. However, identified DE genes that are associated with the fate switch has to be investigated to pick genes for further analysis and perform *in-vitro* or *in-vivo* experiments to reveal their biological role in the VMN-to-5HTN fate switch process. We compared our identified common DE genes with published clusters of genes known to be associated with Shh signaling [32], and the analysis gave some shared genes [Appendix A]. Such kind of analysis helps to pick genes associated with Shh, as it is an important signaling pathway that initiates the entire process in study.

The most relevant and suitable pathway analysis approach was chosen among GSEA and NEA. The results obtained by GSEA Pre-ranked were not statistically valid to rely upon. This approach is not suitable when we want to perform higher comparisons. $\text{GSEA}_{\text{binomial}}$ method allows us to make any number of comparisons but the sensitivity of GSEA is less than NEA. NEA performed on the gene list obtained from cross-comparisons, gave stronger enrichment with signaling pathways involving MAPK, Hedgehog, TGFβ2, Wnt and neurotrophin. There exist some limitations for NEA as follows i) networks used as functional gene sets in NEA are not complete. This limits the scope of the analysis but can be overcome by combining all the related network resources ii) genes i.e. members of FGS often overlap across pathways which makes difficult to highlight on genes that constitute the pathways.

CONCLUSION

This master thesis work is only a part of a long term research focused in studying the mechanisms regulating temporal cell-fate generation. Further *in-vitro/in-vivo* experiments are required to identify and prove the novel candidate genes. Research is ongoing to overcome the problems dealing with the RNA-seq data. The new approach used in this project to identify differential expressed genes and visualize the differential expressed genes from RNA-seq is promising and efficient to obtain the solution. With the help of statistical analysis huge number of differentially expressed genes was narrow down and a significant number of genes associated with the fate switch from VMN-to-5HTN were emphasized. This allows biologists to focus on the short list of genes for further analysis and reduce the labor devoted to experiments. By enrichment analysis signaling pathways such as MAPK, TGFβ2, calcium signaling, Hedgehog were found to be associated to the cell-fate switch process.

6 REFERENCES

1. Pattyn A, Vallstedt A, Dias J, Abdel Samad O, Krumlauf, R., Rijli, F.M, Brunet, J-F, Ericson, J. (2003). Coordinated temporal and spatial control of motor neuron and serotonergic neuron generation from a common pool of CNS progenitors. *Genes Dev 17*, 729-737.
2. Vallstedt, Klos J.M, Ericson J (2005) Multiple dorsoventral origins of oligodendrocyte generation in the spinal cord and hindbrain. *Neuron 45*, 55-67.
3. Muller, C.P and Jacobs, B.L. (2010) Handbook of the Behavioral Neurobiology of Serotinin, *Academic Press*
4. Kanning KC, Kaplan A, Henderson CE. (2010) Motor neuron diversity in development and disease. *Annu Rev Neurosci., 33*, 409-40
5. Kinney, H.C. (2005) Abnormalities of the brainstem serotonergic system in the sudden infant death syndrome: a review. *Pediatr Dev Pathol. 8*, 507-24.
6. Mann, J.J. (1999). Role of the serotonergic system in the pathogenesis of major depression and suicidal behavior. *Neuropsychopharmacology. (2 Suppl):*99S-105S.
7. Sang-Hun L, Nadya L, Lorenz S, Auerbach J.M, McKay R.D (2000) Efficient generation of midbrain and hindbrain neurons from mouse embryonic stem cells, *Nature Vol 18*, 675-679
8. Pierani A, Brenner M.S, Chiang C, Jessell TM (1999) A sonic hedgehog independent, retinoid activated pathway of neurogenesis in the ventral spinal cord , *Cell Vol 97*, 903-915
9. Jacob J, Ferri AL, Milton C, Prin F, Pla P, Lin W, Gavalas A, Ang SL, Briscoe J. (2007) Transcriptional repression coordinates the temporal switch from motor to serotonergic neurogenesis, *Nat Neuroscience 10*, 1433-1439.
10. Jose M. Dias et.al. A Temporal Signal Relay Mechanism by Shh and Tgfβ2 Underlies the Sequential Specification of Motor Neurons and Serotonergic Neurons in the Developing CNS, Manuscript
11. Wang Z, Gerstein M, and Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics *.Nature Reviews Genetics.*10: 57-63.
12. Oshlack A, Robinson M.D, Young M.D. (2010) From RNA-seq reads to differential expression a result. *Genome Biology.*11:220
13. Alexeyenko A, Lee W, Pernemalm M, Guegan J, Dessen P, Lazar V, Lehtiö J, Pawitan Y. (2012) Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics.*13:226.

14. McCormack T, Frings O, Alexeyenko A, Sonnhammer E, (2013) Statistical Assessment of Crosstalk Enrichment between Gene Groups in Biological Networks, *PLOS ONE, 18*: e54945.
15. Benjamini Y, Hochberg Y. (1995) Controlling False Discovery Rate: A practical and powerful approach to multiple testing, *Journal of Royal statistical society, 57:* 289-300.
16. Bland J M and Altman D G. (2000) Statistics Notes: The Odds ratio. *BMJ*. 320: 1468.
17. Alexeyenko A and Sonnhammer E. (2009) Global networks o functional coupling in eukaryotes from comprehensive data integration. *Genome Research, 19*:1107-1116.
18. Stacey D B, Christopher M, Nicholas J. L. and Anita B. R (2004) SB-505124 Is a Selective Inhibitor of Transforming Growth Factor- β Type I Receptors ALK4, ALK5, and ALK7 *MolPharmacol 65*:744-752.
19. http://www.scilifelab.se/archive/tmp/stockholm/20144_03%20Sample%20requirements%20for%20genomics%20projects.pdf
20. TopHat : tophat.cbcb.umd.edu/manual.html
21. HTSeq: Analysis high-throughput Sequencing data with Python: [http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html]
22. Cufflinks : cufflinks.cbcb.umd.edu
23. A.K. Jain, M. M. (1999). Data Clustering: A Review. *ACM Computing Surveys*, 31(3), 264-323.
24. Johnson S.C. (1967) Hierarchical Clustering Schemes , *Psychometrika*, 2:241-254
25. Anders S, Huber W. (2010) Differential Expression analysis for sequence count data. *Genome Biology* :11:R106
26. Subramanian A, Tamayo P, Vamsi K. M, Mukherjee S, Benjami L.E, Paulovich A, Pomeroy S.L, Golub T. R, Eric S. Lander E.S, Mesirov J.P. (2005) Gene Set enrichment analysis: A knowledge-bases approach for interpreting genome-wide expression profiles, *PNAS, 102*:15545-15550.
27. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acid Research, 40,* D109-D114.
28. The Gene Ontology Consortium. (May 2000) Gene ontology: tool for the unification of biology. 25(1):25-9.
29. Ruepp A, Brauner B, Dunger Kaltenbach I, Frishman G, Montrone C, Stransky M, Waegele B, Schmidt T, Doudieu ON, Stümpflen V, Mewes HW. (2008) CORUM: the comprehensive resource of mammalian protein complexes, *Nucleic Acids Research, 36*- D646-50.
30. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker

T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res, 13*: 2498-504.

31. Webpage: http://research.scilifelab.se/andrej_alexeyenko/HyperSet/Ashwini/DE_conditions.html
32. Catarina Cruz, (2010) Foxj1 regulates floor plate cilia architecture and modifies the responses of cells to sonic hedgehog signaling, *Development, 137:4271*-4282.
33. Jose Dias, “Spatial and Temporal Mechanisms of Cell Fate Determination in The Developing CNS”, Karolinska Institute, Report number: 978-91-7457- 457-9, Pages 59.

APPENDIX A

Common genes identified between our dataset and published dataset.

Cluster1: Shh signaling down-regulated genes, Cluster4: Up-regulated genes by Shh signaling

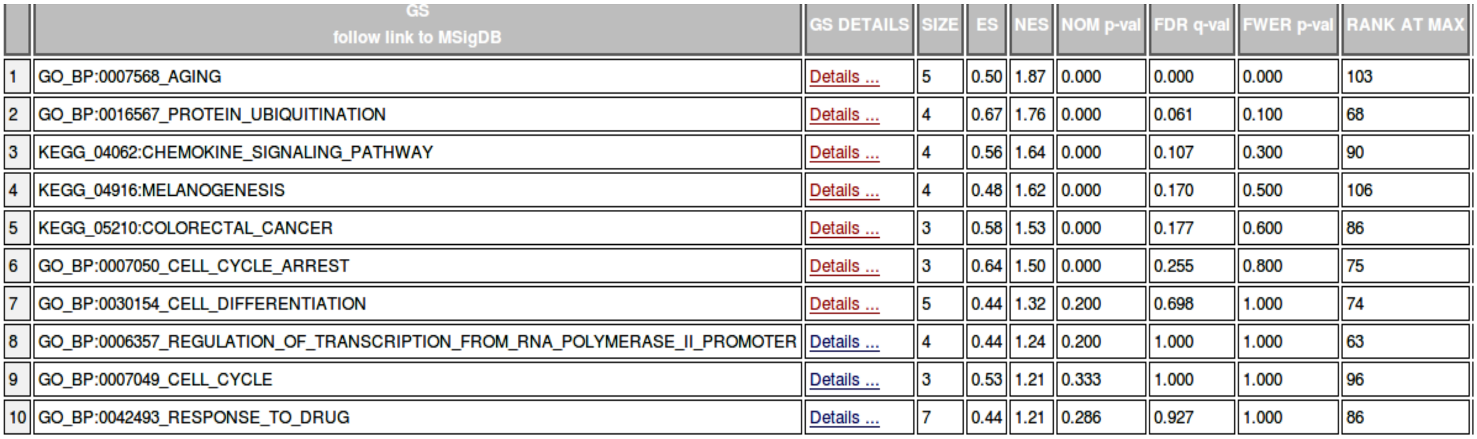
wt 5.5-wt3.5	Up	Cluster 1	4	Enpp2,Ednra,Cbln1,Jam2
		Cluster4	16	Sulf2,Pltp,Slc18a2,Itm2c,Atp1b1, Ablim2,Plekha2,Foxj1,Spon1,Slc43a2,Igfbp2,Ttc9,Sdpr,Lrrc49,Cntn2,Sepp1
	Down	Cluster 1	6	Gli3,Dusp1,Pax6,Akap12,Nexn,Rasl11b
		Cluster 4	11	Bid,Bcl2l1,Fst,Tle4,Anxa1,Bmp2,Cgnl1,Stra6,Cdk6,Cyp26b1,Lgals1
wt3.5-Phox3.5	Up	Cluster1	1	Jam2
		Cluster 4	1	Ablim2
	Down	Cluster 1	1	Pax6
		Cluster 4	2	Cyp26b1,Cgnl1
wt5.5-SB5.5	Down	Cluster1	1	Akap12

Studies on Affymetrix Genechip in chick neural progenitors, the Shh signaling had been cell-autonomously activated/blocked resulting in the identification of two interesting clusters of genes. Cluster 1 comprises of genes down-regulated by Shh signaling and other cluster 4 with genes that induced by Shh [32].

APPENDIX B

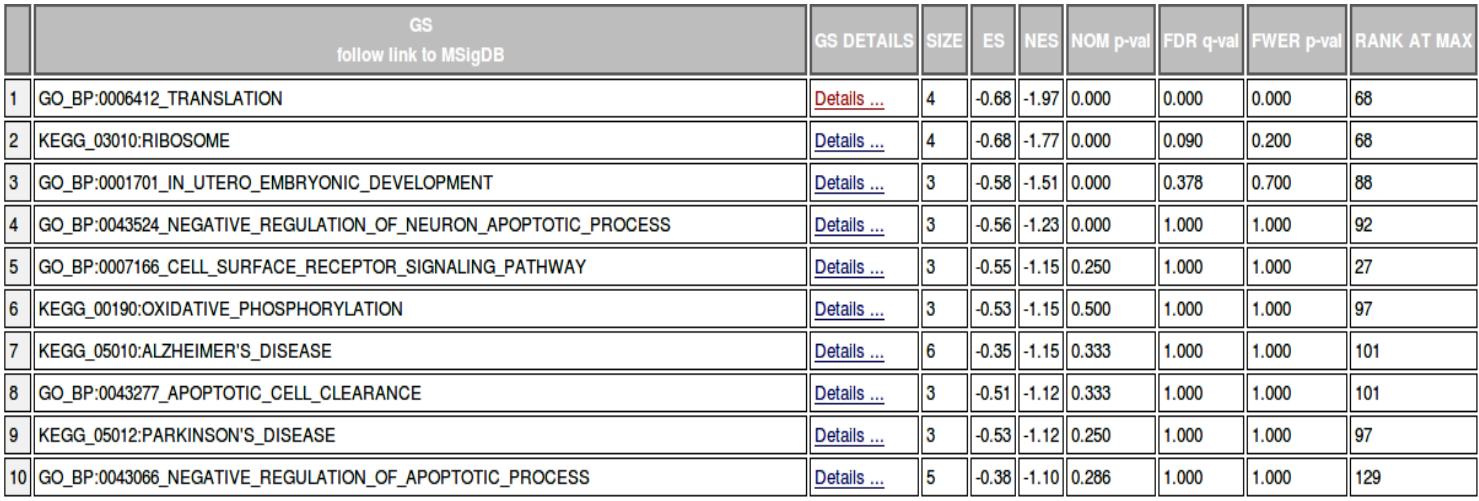
GSEA-Preranked tool output

List of the first 10 over-represented gene sets that have positive enrichment scores (top of the ranked list)



	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
1	GO_BP:0007568_AGING	Details ...	5	0.50	1.87	0.000	0.000	0.000	103
2	GO_BP:0016567_PROTEIN_UBIQUITINATION	Details ...	4	0.67	1.76	0.000	0.061	0.100	68
3	KEGG_04062:CHEMOKINE_SIGNALING_PATHWAY	Details ...	4	0.56	1.64	0.000	0.107	0.300	90
4	KEGG_04916:MELANOGENESIS	Details ...	4	0.48	1.62	0.000	0.170	0.500	106
5	KEGG_05210:COLORECTAL_CANCER	Details ...	3	0.58	1.53	0.000	0.177	0.600	86
6	GO_BP:0007050_CELL_CYCLE_ARREST	Details ...	3	0.64	1.50	0.000	0.255	0.800	75
7	GO_BP:0030154_CELL_DIFFERENTIATION	Details ...	5	0.44	1.32	0.200	0.698	1.000	74
8	GO_BP:0006357_REGULATION_OF_TRANSCRIPTION_FROM_RNA_POLYMERASE_II_PROMOTER	Details ...	4	0.44	1.24	0.200	1.000	1.000	63
9	GO_BP:0007049_CELL_CYCLE	Details ...	3	0.53	1.21	0.333	1.000	1.000	96
10	GO_BP:0042493_RESPONSE_TO_DRUG	Details ...	7	0.44	1.21	0.286	0.927	1.000	86

The list of over-represented genes identified by GSEA-Preranked tool, showing negative enrichment scores (bottom of the ranked list)



	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
1	GO_BP:0006412_TRANSLATION	Details ...	4	-0.68	-1.97	0.000	0.000	0.000	68
2	KEGG_03010:RIBOSOME	Details ...	4	-0.68	-1.77	0.000	0.090	0.200	68
3	GO_BP:0001701_IN_UTERO_EMBRYONIC_DEVELOPMENT	Details ...	3	-0.58	-1.51	0.000	0.378	0.700	88
4	GO_BP:0043524_NEGATIVE_REGULATION_OF_NEURON_APOPTOTIC_PROCESS	Details ...	3	-0.56	-1.23	0.000	1.000	1.000	92
5	GO_BP:0007166_CELL_SURFACE_RECEPTOR_SIGNALING_PATHWAY	Details ...	3	-0.55	-1.15	0.250	1.000	1.000	27
6	KEGG_00190:OXIDATIVE_PHOSPHORYLATION	Details ...	3	-0.53	-1.15	0.500	1.000	1.000	97
7	KEGG_05010:ALZHEIMER'S_DISEASE	Details ...	6	-0.35	-1.15	0.333	1.000	1.000	101
8	GO_BP:0043277_APOPTOTIC_CELL_CLEARANCE	Details ...	3	-0.51	-1.12	0.333	1.000	1.000	101
9	KEGG_05012:PARKINSON'S_DISEASE	Details ...	3	-0.53	-1.12	0.250	1.000	1.000	97
10	GO_BP:0043066_NEGATIVE_REGULATION_OF_APOPTOTIC_PROCESS	Details ...	5	-0.38	-1.10	0.286	1.000	1.000	129

APPENDIX C

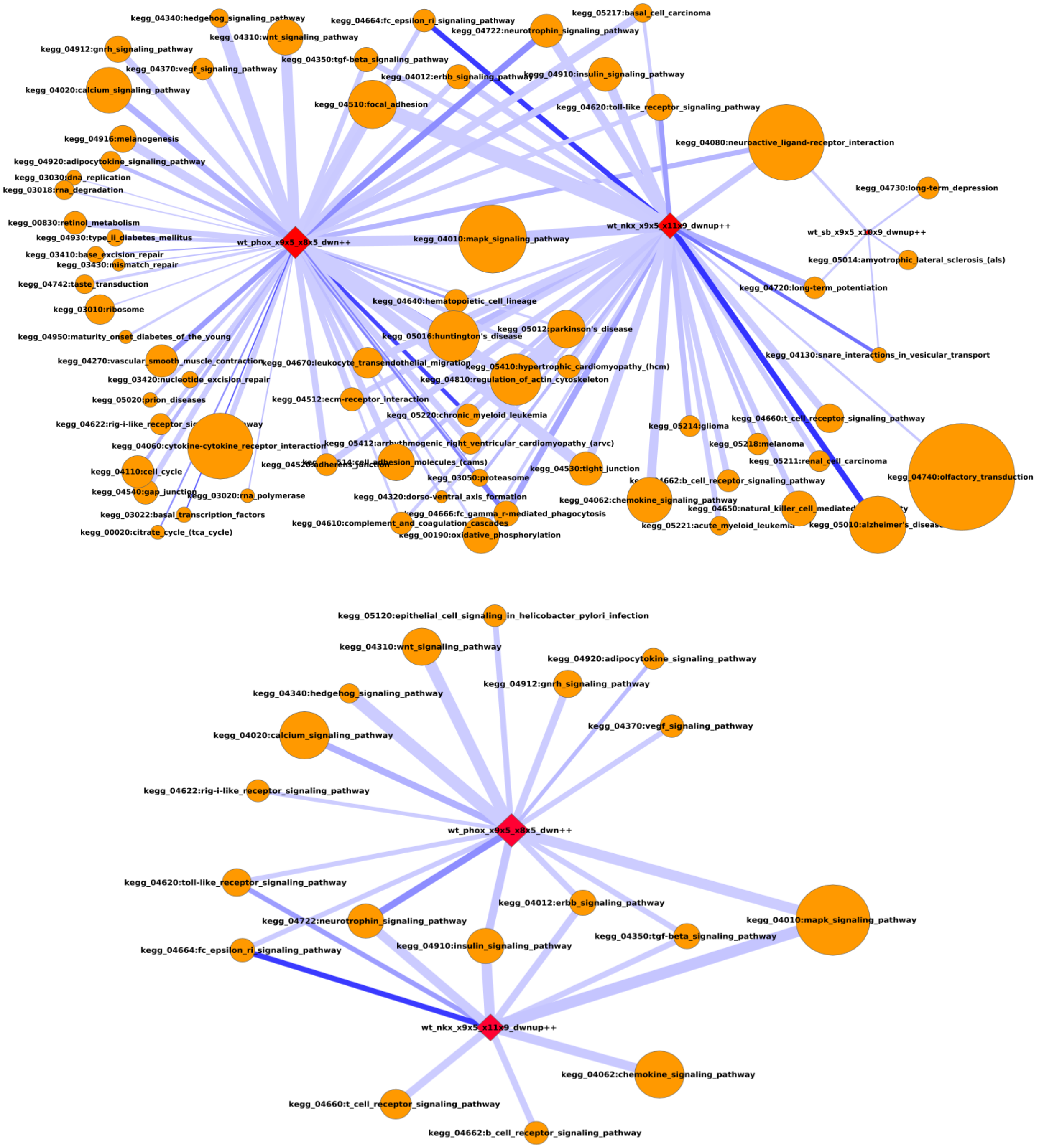
Quadrapulet comparisons between WT and mutant samples

Condition	Quadruplet Comparison	One-Fold	Two-Fold
wild type Vs. Phox2b	***x5x8,x5x9,x8x12,x9x12***	**2782**	**1024**
Nkx2.2 Vs. Phox2b	***x7x8,x7x11,xx8x12,x11x12***	**2516**	**894**
Nkx2.2 Vs. SB	***x6x7,x6x10,x7x11,x10x11***	**2019**	**779**
wild type Vs. Nkx2.2	***x5x7,x5x9,x7x11,x9x11***	**1983**	**751**
wild type Vs. SB	***x5x6,x5x9,x6x10,x9x10***	**1644**	**604**

Total number of significant differentially expressed genes identified by the quadruplet comparisons

APPENDIX D

Cytoscape visualization of NEA analysis for 103 DE genes associated with VMN generation


kegg_04340:hedgehog_signaling_pathway
kegg_04310:wnt_signaling_pathway
kegg_04664:fc_epsilon_ri_signaling_pathway
kegg_04722:neurotrophin_signaling_pathway
kegg_05217:basal_cell_carcinoma
kegg_04912:gnrh_signaling_pathway
kegg_04370:vegf_signaling_pathway
kegg_04350:tgf-beta_signaling_pathway
kegg_04012:erbb_signaling_pathway
kegg_04910:insulin_signaling_pathway
kegg_04020:calcium_signaling_pathway
kegg_04510:focal_adhesion
kegg_04620:toll-like_receptor_signaling_pathway
kegg_04916:melanogenesis
kegg_04080:neuroactive_ligand-receptor_interaction
kegg_04920:adipocytokine_signaling_pathway
kegg_03030:dna_replication
kegg_03018:rna_degradation
kegg_04730:long-term_depression
kegg_00830:retinol_metabolism
kegg_04930:type_ii_diabetes_mellitus
wt_phox_x9x5_x8x5_dwn++
kegg_04010:mapk_signaling_pathway
wt_nkx_x9x5_x11x9_dwnup++
wt_sb_x9x5_x10x9_dwnup++
kegg_03410:base_excision_repair
kegg_03430:mismatch_repair
kegg_05014:amyotrophic_lateral_sclerosis_(als)
kegg_04742:taste_transduction
kegg_04720:long-term_potentiation
kegg_04640:hematopoietic_cell_lineage
kegg_03010:ribosome
kegg_05012:parkinson's_disease
kegg_04950:maturity_onset_diabetes_of_the_young
kegg_05016:huntington's_disease
kegg_04130:snare_interactions_in_vesicular_transport
kegg_04270:vascular_smooth_muscle_contraction
kegg_04670:leukocyte_transendothelial_migration
kegg_05410:hypertrophic_cardiomyopathy_(hcm)
kegg_03420:nucleotide_excision_repair
kegg_04810:regulation_of_actin_cytoskeleton
kegg_05020:prion_diseases
kegg_04512:ecm-receptor_interaction
kegg_05220:chronic_myeloid_leukemia
kegg_04622:rig-i-like_receptor_si way
kegg_04660:t_cell_receptor_signaling_pathway
kegg_05214:glioma
kegg_04060:cytokine-cytokine_receptor_interaction
kegg_05412:arrhythmogenic_right_ventricular_cardiomyopathy_(arvc)
kegg_05218:melanoma
kegg_04520:adherens_junction
ke 514:cell_adhesion_molecules_(cams)
kegg_05211:renal_cell_carcinoma
kegg_04110:cell_cycle
kegg_04530:tight_junction
kegg_03050:proteasome
ke 4662:b_cell_receptor_signaling_pathway
kegg_04740:olfactory_transduction
kegg_04540:gap_junction
kegg_03020:rna_polymerase
kegg_04320:dorso-ventral_axis_formation
kegg_04062:chemokine_signaling_pathway
kegg_04650:natural_killer_cell_mediated ty
kegg_03022:basal_transcription_factors
kegg_04666:fc_gamma_r-mediated_phagocytosis
kegg_04610:complement_and_coagulation_cascades
kegg_05221:acute_myeloid_leukemia
kegg_05010:alzheimer's_diseas
kegg_00020:citrate_cycle_(tca_cycle)
kegg_00190:oxidative_phosphorylation
kegg_05120:epithelial_cell_signaling_in_helicobacter_pylori_infection
kegg_04310:wnt_signaling_pathway
kegg_04920:adipocytokine_signaling_pathway
kegg_04340:hedgehog_signaling_pathway
kegg_04912:gnrh_signaling_pathway
kegg_04370:vegf_signaling_pathway
kegg_04020:calcium_signaling_pathway
kegg_04622:rig-i-like_receptor_signaling_pathway
wt_phox_x9x5_x8x5_dwn++
kegg_04620:toll-like_receptor_signaling_pathway
kegg_04012:erbb_signaling_pathway
kegg_04722:neurotrophin_signaling_pathway
kegg_04010:mapk_signaling_pathway
kegg_04350:tgf-beta_signaling_pathway
kegg_04910:insulin_signaling_pathway
kegg_04664:fc_epsilon_ri_signaling_pathway
wt_nkx_x9x5_x11x9_dwnup++
kegg_04062:chemokine_signaling_pathway
kegg_04660:t_cell_receptor_signaling_pathway
kegg_04662:b_cell_receptor_signaling_pathway